Two of a kind?

Vriens, Ingrid

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

TWO OF A KIND?

Comparing Ratings and Rankings for Measuring Work Values using Latent Class Modeling

Ingrid Vriens

TWO OF A KIND?

Comparing Ratings and Rankings for Measuring Work Values using Latent Class Modeling

ISBN: 978-94-6295-295-9

Cover design: Robert Vriens

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Published by: Uitgeverij BOXPress, ‘s-Hertogenbosch

This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), grant number: 400-09-373.

TWO OF A KIND?

Comparing Ratings and Rankings for Measuring Work Values using Latent Class Modeling

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E. H. L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

vrijdag 20 november 2015 om 14.15 uur

door

Ingrid Petronella Maria Vriens

geboren op 28 september 1987 te Deurne

Promotiecommissie:

Promotor:	Prof. dr. J. K. Vermunt
Copromotores:	Dr. J. P. T. M. Gelissen
	Dr. G. B. D. Moors
Overige leden:	Prof. dr. J. W. M. Das
	Prof. dr. H. van Herk
	Prof. dr. E. D. de Leeuw
	Prof. dr. B. Meuleman
	Dr. L. C. J. M. Halman

Contents

CHAPTER 1

Introduction

A key topic in sociological research concerns the study of human values. All major current sociological survey investigations – the European Values Study, the World Values Study, the European Social Survey, and the International Social Survey Project –, seek to empirically measure what people find important in life. In these surveys, researchers predominantly use the rating approach and less often use a ranking method to measure particular value-orientations. In the rating approach respondents are being asked to rate each of the items on a predefined scale (like, for example, a 5-point Likert scale ranging from "very unimportant" to "very important"), while in the ranking approach respondents are being asked to rank-order a number of items based on the importance the respondent attaches to each of the items relative to the other items presented. An important reason why the rating approach is more popular than the ranking approach is that rating data, in contrast to ranking data, can be straightforwardly analyzed with statistical methods with which researchers are familiar. Apart from this, both ratings and rankings for measuring personal values have their own theoretical underpinnings and methodological intricacies and this has led to a body of literature in which the survey-methodological aspects of the rating approach and the ranking approach have been scrutinized. Comparisons of the results of the administration of rating and ranking procedures by existing survey-methodological studies have shown that the results obtained by each approach differ from another (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; Maio, Roese, Seligman, & Katz, 1996; McCarty & Shrum, 2000; Ovadia, 2004). The impetus for

these studies – also for the current study – is the idea that both formats should actually not lead to fundamentally different substantive conclusions concerning the validity of the measurements if particular features of each method are taken into account. This idea came to be known as Krosnick and Alwin's "form-resistant correlation hypothesis" (1985, 1988). The conjecture of the form-resistance correlation hypothesis is important as it directs our attention to the issue that a difference in the results from both approaches may be a consequence of the way a theoretical concept is measured – with the rating or with the ranking method. Then, method-specific features and biases of each response format can have an undesirable systematic influence on the answers given by respondents and the results obtained and thus a researcher should control for these method-specific effects (Alwin & Krosnick, 1985).

The purpose of this study is to bring new empirical evidence to the discussion about the superiority of either ratings or rankings that resulted from the seminal work by Krosnick and Alwin. Specifically, in this dissertation we seek to answer the following general research question: how comparable and consistent are measurements of personal values that result from the application of the rating and ranking approach when we account for method-specific features of each response format? To answer this question, we have developed a survey-experiment that allows not only a between-subjects comparison of the results of ratings and rankings, but also a within-subjects comparison. This design will be described in more detail later on in this introductory chapter. We make use of recent developments in the modeling of rating and ranking data as well as in the modeling of response biases in such measurements. In particular, we apply an innovative modeling approach using latent class modeling to transform the rating data acquired from this design into relative preferences to allow a more systematic comparison of the results of rating data with the results of ranking data. Another novel feature of our study is that with the within-subjects design we are able to investigate the consistency of results of measurement models that are specifically geared towards the analysis

of data resulting from the application of either approach. Specifically, we look at what happens if the same respondent completes twice the rating questionnaire, or twice the ranking task, or first a rating and then a ranking, or vice versa. In all conditions of this design, the same personal values are purported to being measured.

1.1 Theoretical Differences between Ratings and Rankings

An important difference between the ranking and rating approach lies in what is actually being measured with each response format; this issue is closely linked to the question what values are. There are two theoretically different views which form the basis of the ranking versus rating discrepancy. On the one hand there are those who argue that values are hierarchically ordered (Rokeach, 1973, p. p. 5; Schwartz & Bilsky, 1987). This means that a respondent prefers certain values in comparison to other values and that a preference choice can be made even between values that are closely related. According to this view a ranking approach is the most appropriate for the measurement of values. On the other hand there are those who contend that the importance of one value does not necessarily have an effect on other values for an individual and that it should also be possible to assign similar importance scores to values that a respondent finds equally important (Parsons & Shils, 1962, p. p. 405). According to this second view of what values are, the rating approach would be the best method to use. In summary, based on theoretical argument neither approach can be seen as superior for the measurement of values, because it is impossible to know whether values are actually hierarchically ordered or not. Therefore methodological differences also play a crucial role in the discussion on the superiority of either method, which will be discussed below. But first we elaborate in the next section on the details of the research design that we have implemented for the purpose of this study. This will help the reader to get a better grasp

of the empirical part of this study in relation to the survey-methodological research questions that are central to this dissertation.

1.2 Design of the Between-Subjects and Within-Subjects Study

For the investigation of the comparability and consistency of the rating and ranking approach we administered both approaches at two measurement occasions in the LISS (Longitudinal Internet Studies for the Social Sciences) panel of CentERdata. The advantage of using this web panel is that we are able to collect longitudinal data in a large group of respondents which is representative of the Dutch population. The large group of respondents is especially necessary since, as we will see below, we have 10 different conditions (all different combinations of rating and ranking as well as two different orderings of the items within the rating and ranking questionnaires) to which a respondent can belong and we want as many respondents per condition as we possibly can have. Also, since panel members participate monthly in questionnaires this makes it possible not only to investigate the comparability of the rating and ranking method between different individuals, but also to investigate how consistent respondents are in answering the question about human values, given the experimental condition to which they belong. The rating and ranking questionnaires that have been used throughout all the chapters of this dissertation were collected in a small survey-experiment with the first measurement in June and July 2012 and the second measurement in September and October 2012. In particular, the design we use is a split-ballot design with repeated measures. We start with a between-subjects design which is extended with repeated measures into a within-subjects design. Only respondents that filled in the questionnaire at the first measurement occasion also received the second questionnaire. In the between-subjects design we implement two different versions (A and B) of 17 rating and ranking questions; see Table 1.1 for details.

Table 1.1 Questionnaire design

Ordering of job aspect items in two experimental conditions		
	Version A	**Version B**
(1)	Good pay	(9)
(2)	Pleasant people to work with	(8)
(3)	Not too much pressure	(7)
(4)	Good job security	(6)
(5)	Good hours	(5)
(6)	An opportunity to use initiative	(4)
(7)	A useful job for society	(3)
(8)	Generous holidays	(2)
(9)	Meeting people	(1)
(10)	A job in which you feel you can achieve something	(17)
(11)	A responsible job	(16)
(12)	A job that is interesting	(15)
(13)	A job that meets one´s abilities	(14)
(14)	Learning new skills	(13)
(15)	Family friendly	(12)
(16)	Have a say in important decisions	(11)
(17)	People treated equally at the workplace	(10)
Question format: ranking		
(a)	Here are some aspects of a job that people say are important. The question is which of these you personally think is the most important in a job?	
(b)	Of the remaining aspects of a job, which one do you consider next most important?	
(c)	Of the remaining aspects of a job, which one do you then consider next most important?	
(d)	And which one of the remaining aspects do you consider least important of all?	
Question format: rating		
	Here are some aspects of a job that people say are important: How important is each of these to you personally?	
	1 Very unimportant 2 3 4 5 Very important	

The questionnaires were about well-known and much investigated types of personal values, namely work values – what people find important in a job. In particular, the questionnaires consisted of 17 different job aspects that people could find important. In the rating task, we asked respondents to indicate how important each of the job aspects were to them personally on a 5-point scale. In the ranking task we asked respondents to rank their top 3 most important items and the least important one out of the full item list. The difference between versions A and B is that the order in which the work values items are presented is changed in order to investigate order effects. See the last column of the upper part of Table 1.1 for the change in ordering of the items being presented to respondents.

In a next stage of the study, the initial between-subjects design is extended into a within-subjects design by implementing a repeated measurement in which the initial four conditions are further randomly subdivided (see, for the blueprint details of the design, Table 1.2). In the repeated measurement we distinguish ten conditions, four of which are control conditions which allow us to investigate the stability of the responses to questions which were of the same question format and version (rating or ranking, A or B). In the remaining conditions, both rating and ranking are offered interchangeably to the respondents.

Table 1.2 Split ballot design with repeated measurements

Condition		T_1	T_2
1	R	$O_{Rating.Version\ A}$ (N=2000)	$O_{Rating.Version\ A(N=500)}$ $O_{Ranking.\ Version\ A\ (N=500)}$ $O_{Rating.Version\ B\ (N=500)}$ $O_{Ranking.\ Version\ B\ (N=500)}$
2	R	$O_{Rating.Version\ B}$ (N=400)	$O_{Rating.Version\ B\ (N=400)}$
3	R	$O_{Ranking.Version\ A}$ (N=2000)	$O_{Ranking.Version\ A\ (N=500)}$ $O_{Rating.\ Version\ A\ (N=500)}$ $O_{Ranking.\ Version\ B\ (N=500)}$ $O_{Rating.\ Version\ B\ (N=500)}$
4	R	$O_{Ranking.Version\ B}$ (N=400)	$O_{Ranking.Version\ B\ (N=400)}$

1.3 Ratings, Rankings and Method-Specific Response Biases

Since the only methodological difference between the rating and ranking approach is the way the alternatives are shown to the respondents (under the assumption that there is no difference in the content of the question or items/alternatives), ideally the results obtained by using either approach should be similar. However, answers given to the questions framed in either format may be affected by response biases. The term response bias refers to "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Baumgartner & Steenkamp, 2001; Paulhus, 1991, p. p. 17). Response biases lead to answers that do not only reflect the substantive meaning attached to a question but also the tendency of a respondent to respond in a certain manner. Because of the difference in the answering task the effect of response bias will be different for rating items and ranking items.

A type of response bias that we expect to particularly influence the rating results is the overall level of agreement or importance. In the current study we are interested in what people find important in a job and as one can imagine this question may lead to answers that indicate that all of the items or alternatives presented to the respondents are being judged as important. This would mean that respondents would only use half of the scale presented to them and that there is not much variation left in the answers provided by respondents. Since the absolute level of importance would not display much variation, we argue it will be more informative to look at the relative preferences for each of the items. To this end, we apply statistical models that allow us to 'rank the ratings' while controlling for the overall level of agreement in the rating answers given by respondents.

Another response bias that is expected to be present in the current rating data is the non-differentiation response style or 'straight-lining'. Respondents being susceptible to this particular response style have the tendency to not really differentiate between the ratings they

give to each of the items by giving identical (or nearly identical) responses to all items. Of course it is possible that this rating scheme follows from a critical consideration of each of the alternatives; however, it can also be the consequence of satisficing behavior. When a respondent is satisficing instead of optimizing, it means that the respondent is less motivated to make a cognitive effort to provide an optimal answer to each of the items; instead, he or she is leaning towards giving an easier answer like staying with the first point of the rating scale that was selected (Krosnick, 1991). In this study the latent class modeling approach that we use makes it possible to identify a group of non-differentiating respondents. By identifying this specific group or respondents, the results for the remaining respondents will not be confounded by non-differentiation.

The response biases that may have an effect when using a rating scale can actually be overcome by using the ranking approach. In the ranking approach respondents are forced to discriminate between the items given to them and this approach avoids decision-making about the numbering or labeling of the rating scale (DeCarlo & Luthar, 2000). However, there is a response bias that has often been found to affect ranking data, namely the response order effect (Becker, 1954; Campbell & Mohr, 1950; Fuchs, 2005; Klein, Dulmer, Ohr, Quandt, & Rosar, 2004; Krosnick, 1992; Krosnick & Schuman, 1988; McClendon, 1986; McClendon, 1991; Schuman & Presser, 1996; Stern, Dillman, & Smyth, 2007). The response order effect means that an item has a higher probability of being chosen as one of the most important items just because of its placement in the full list of items (i.e. at the top or the end of the list) and not because of the content of the item. In particular, if some respondents are predisposed to selecting the first response options, this is a primacy effect, and if they are predisposed to selecting the last response options, this is called a recency effect.

Usually a response order effect is controlled in a given sample by showing the items to the respondents in a fully randomized order. However, this approach may not always be

feasible since it requires a larger sample size as the size of the item set increases. In this dissertation we develop a novel and more efficient approach of statistically controlling for a response order effect that may be present in the ranking data. With this new method only two different orderings of the items are needed to be able to get an estimation of the response order effect. This analysis is based on the between-subjects part of our research design.

1.4 Consistency of Measurements with Ratings and Rankings

In addition to comparing the rating and ranking methods between different individuals, we are also interested in the comparison of both approaches within the same individuals. This issue is investigated using the within-subjects design element in our survey experiment. First of all, this design gives more insight into how stable the results of both the rating and the ranking approaches are. Ideally, the results between measurement occasions should not be very different since the only thing that differs between the rating and ranking approach is the way the respondents have to formulate their answer. Secondly, and more interestingly, we investigate how respondents who on a first measurement moment have responded in a rating format respond on a repeated measurement in a ranking format, or vice versa. Specifically, the within-subjects design allows not only for the investigation of how stable the content-related value measurements are, but also how stable response biases are over time. Based on this information we may infer whether response biases are some sort of person-dependent trait in which certain respondents always have the tendency to show this behavior or whether it may be a tendency which occurs occasionally. Related to this it is interesting to look at what happens if respondents who portray a response bias (e.g., non-differentiation) at one measurement occasion are investigated again by confronting them the second time around

with a different response format that logically excludes the possibility of portraying the response tendency.

1.5 Comparability of Rating and Ranking Approaches while Controlling for Response Biases: The Latent Class Modeling Approach

As stated above the comparison of ratings and rankings for the measurement of values is not new. However, in previous research conclusions about the comparability of the two methods are based on either descriptive analyses or factor-analytical procedures for continuous-level variables. In the current study a latent class modeling approach is being used, which has several benefits compared to the methods used in previous research. First, in contrast to traditional factor analysis in which both the observed variables (ratings and rankings in our case) and the latent variables of interest are being treated as if they follow a continuous measurement scale, the latent class modeling approach does not need such assumptions. Second, the latent class approach is able to model the actual choice process of the ranking data, which overcomes the problem of ipsativity of the data (which means that the sum of items is the same for all respondents as a result of the dependency that exists between choices). A set of constraints is needed to correct for this ipsativity when a factor analytic approach is used, but the latent class modeling approach makes it possible to analyze the ranking data without having to make any adjustments. Third, the model that we use allows for the estimation of – and controlling for – response biases in the data while at the same time the concept of interest (in the current study personal work values preferences) is measured. For rating data we also use the latent class modeling approach. Here, the previously mentioned overall agreement tendency in rating data is controlled for by applying a model-based transformation of the rating items which makes it possible to distinguish the overall

agreement or importance from the relative preferences of each of the items. If non-differentiating respondents are present in the rating data, these respondents will become visible as a separate latent segment with item preference values that do not differ from zero much. Response order effects (if present) in the ranking data are accounted for by including a choice-specific attribute in the latent class model. This choice attribute will influence the rank of the choices made by respondents, but only for the items that are shown first or last in the list of items. All these benefits of the latent class modeling approach make it a very useful tool for the comparison of ratings and rankings.

1.6 Outline of the Dissertation

The chapters in this dissertation are all related to the comparison of rating and ranking methods for the measurement of values, while accounting for method-specific features. Since this dissertation is written in such a way that each of the chapters can be read independently of the remaining text, it is impossible to avoid repetition when explaining central concepts of the dissertation. Below, a short overview of each of the chapters is being presented.

Chapter 2

In this chapter a method is being presented to control for response order effects in ranking data. The ranking task that respondents received was to rank the top 3 most important and the least important item out of a set containing 17 items concerning work values. We use two different orderings of the items to be able to investigate the response order effect. In the current data we find evidence of the existence of a primacy effect. Using the Latent Choice Factor (LCF) model we are able to analyze the response order effects in a straightforward

way. Comparing the model with control for response order effect with the model without a control for response order effect shows that the rank order of alternatives changed. Finally, with this modeling approach we are able to acquire an estimate of the size of the primacy effect.

Chapter 3

This chapter presents a latent class segmentation approach for the between-subjects comparison of rating and ranking procedures in which it is also possible to take into account the method-specific features. The method-specific features which are being controlled in this study are the response order effect in the ranking questionnaire (using a similar approach as described in chapter 2) and overall agreement and non-differentiation in the rating task. To make the two response formats more comparable and to control for the overall level of agreement or importance, the rating data are transformed into relative preference data. Using the latent class segmentation approach we are able to distinguish two segments with similar item preference structures, regardless of whether the ranking or rating response format is being used. Also, we find segments that differ in preference structure between the two approaches. One of the segments specific for the rating approach consists of non-differentiating respondents. Besides the comparison of the item preference structures, we also investigate the relationships of the latent segments with external variables (such as age or gender) and based on this investigation we find resemblances between the covariate effects and the two similar latent segments in the rating and ranking approach. Thus it is shown that the latent class segmentation approach is a valuable tool for the comparison of ratings and rankings while at the same time being able to account for response biases present in the data.

Chapter 4

In this chapter not only the comparability of the rating and ranking approach for the measurement of values within-subjects is being investigated, but also how consistent the results are over time. Respondents received a questionnaire twice with at least two months in between the two measurement occasions. All possible combinations were investigated: respondents could receive either the rating or the ranking questionnaire twice or a combination of both. Questions of interest are: 1) How consistent are respondents when receiving the same questionnaire with the same method twice? 2) To what extent are the answers given to the rating questionnaire mirrored by using the ranking approach or vice versa? 3) What happens with the respondents belonging to the non-differentiation group in the rating response format when these respondents receive the ranking questionnaire? To link the two measurement approaches at the two measurement occasions, we investigate the cross-classifications between latent segments with similar meanings based on the item preference structure. Also, we test for measurement equivalence in the case that respondents received the same measurement method twice. We find that respondents classified into the intrinsic and extrinsic classes are consistently classified as such across measurement occasions, irrespective of the measurement method being used. Other method-specific latent classes are found to be consistent over time (when the same measurement method was being used). The respondents belonging to the non-differentiation group in the rating questionnaire are found to be equally spread over the latent segments when the ranking approach is used. The measurement equivalence models that are fitted to the ranking twice and rating twice conditions show an improvement in model fit and more pronounced associations between the repeated measurements.

Chapter 5

In this chapter, we summarize the main findings of our study in the light of the research questions. In addition, we discuss the findings within the broader discussion about the 'superiority' of either ratings or rankings. Finally, we discuss some limitations of our study and propose some suggestions for further survey-methodological research on the comparison between ratings and rankings.

CHAPTER 2

Controlling for Response Order Effects in Ranking Items Using Latent Choice Factor Modeling*

Abstract

Measuring values in sociological research sometimes involves the use of ranking data. A disadvantage of a ranking assignment is that the order in which the items are presented might influence the choice preferences of respondents regardless of the content being measured. The standard procedure to rule out such effects is to randomize the order of items across respondents. However, implementing this design may be impractical and the biasing impact of a response order effect cannot be evaluated. We use a latent choice factor (LCF) model that allows statistically controlling for response order effects. Furthermore, the model adequately deals with the known issue of ipsativity of ranking data. Applying this model to a Dutch survey on work values, we show that a primacy effect accounts for response order bias in item preferences. Our findings demonstrate the usefulness of the LCF modeling ranking data while taking into account particular response biases.

* This chapter is accepted for publication as: Vriens, I., Moors, G., Gelissen, J. & Vermunt, J. K. (in press). Controlling for response order effects in ranking items using latent choice factor modeling. *Sociological Methods & Research.*

2.1 Introduction

The most often-used method for measuring values in social surveys is the rating approach in which respondents are asked to indicate their level of agreement with a set of items. However, researchers may prefer another approach for theoretical or methodological reasons: the ranking task, in which respondents are asked to rank-order a limited number of items (which are the response alternatives from which one can choose) according to the respondent's attributed importance to some (partial ranking) or all (full ranking) items. The theoretical impetus for using the ranking method can be found in Rokeach's conceptualization that "a value is an enduring belief that a specific mode of conduct or end-state of existence is personally preferable to an opposite or converse mode" (Rokeach, 1973). Such a systematic preference order is best empirically measured using rank-order scaling. Similarly, Schwartz and Bilsky (1987) identify the ordering of concepts or beliefs that make up values by their relative importance as one of the key characteristics of measuring values.

There are also important methodological reasons for using the ranking procedure because it overcomes several limitations of the rating procedure. First, it forces respondents to make a choice between the given response alternatives, which presumably leads to more informative data (DeCarlo & Luthar, 2000; Maio, Roese, Seligman, & Katz, 1996; Ovadia, 2004) although overall levels of importance cannot be assessed. Second, it avoids the arbitrary decisions that respondents make – for example when using a Likert rating scale – conditional on the number and labels of response options (DeCarlo & Luthar, 2000). Third, ranking data are not affected by response biases like extreme response style (the tendency to choose the highest or lowest possible rating to each item, irrespective of item content) or agreement bias (tendency to agree with all items irrespective of item content). It is well-known that such systematic measurement errors can seriously bias empirical findings (Alwin & Krosnick, 1985).

Despite of these advantages, the use of the ranking method itself is not very popular. An important reason for this are some statistical properties of ranking data (which will be explained later on) that prohibit the straightforward use of statistical methods with which researchers are familiar. A particularly strong point of the modeling approach used in this study is that it adequately deals with the statistical problems associated with ranking data. However, the ranking assignment as such may equally produce specific problems. First, respondents interpret them as more difficult because rankings require a high level of cognitive effort. This difficulty increases as the list of response alternatives gets longer (Alwin & Krosnick, 1985; Galesic, Tourangeau, Couper, & Conrad, 2008; McCarty & Shrum, 2000). Another well-known cause for bias is the order in which the response alternatives are presented to respondents. Previous research has shown that a difference in response order can lead to very different results (Becker, 1954; Campbell & Mohr, 1950; Fuchs, 2005; Klein, Dulmer, Ohr, Quandt, & Rosar, 2004; Krosnick, 1992; Krosnick & Schuman, 1988; McClendon, 1986; McClendon, 1991; Schuman & Presser, 1996; Stern, Dillman, & Smyth, 2007). Specifically, response alternatives shown first or last have a higher probability of being selected just because of their placing in the full list of alternatives and not because of their meaning, which leads to a primacy effect in the former case and to a recency effect in the latter. Krosnick (2000) has presented an overview of previous response order studies that makes clear that visually presented questions elicit primacy effects, while for orally presented questions the recency effect is more present. An important explanation for the occurrence of the response order effect that has been offered in the literature is satisficing behavior (Krosnick & Alwin, 1987; Krosnick & Presser, 2010; Siminski, 2008). Krosnick and Alwin (1987) defined satisficing as “instead of looking for an optimal solution, going for the acceptable option”. In other words, respondents who show satisficing behavior have the

tendency to choose the first alternative that seems to be a reasonable choice to them instead of choosing the most appropriate alternative.

Given the existing evidence regarding response order effects in ranking data statistical models for such data need to adjust for this source of bias. The way this response bias is usually controlled is by randomizing the order in which the items are being shown to respondents. As such it is assumed that any response order effect is ruled out by this procedure. When applied correctly the randomized ordering of items leads to unbiased estimates of group comparisons. However, at the individual level biases remain and – even more problematic – we have no way of accounting for the impact of the response order bias. Furthermore, from a practical point of view it might not always be possible to use a randomized design, for instance with self-administered questionnaires or when show cards are used to facilitate the respondent's task. Another important limitation of the randomization approach is that it requires a relatively large sample to be efficient in reducing random error caused by response order effects. Finally, Dillman and Christian (2005) warn that the randomization approach could have undesirable consequences when measuring change between data collections.

In this paper we present an innovative approach to statistically control for the response order effect by explicitly taking this effect into account in a latent variable model for measuring a substantive or content factor. Not only does our model allow investigating a response order effect, it also enables the researcher to test whether this effect is caused by – for instance – primacy and/or recency effects. Primacy refers to an increased preference for items listed in the beginning of the set whereas recency implies higher preferences for items listed last, regardless of the content of the items presented first or last. An additional benefit of our approach is that a research design with a completely randomized ordering of response alternatives that are shown to respondents is not necessary. Rather, a considerably less

complex split-ballot design with a limited number of conditions in which the order of response alternatives is systematically varied and randomly assigned to respondents is sufficient to implement the modeling approach. A final advantage of the proposed approach is that by controlling for a response order effect we can gain more knowledge about the relative bias of the presentation order of response alternatives on the actual model parameters.

In this study we make use of recent developments in the field of latent class analysis that allow us to define a measurement model that, first, overcomes the inherent statistical issues of modeling ranking data and, second, that allows us to derive an empirical estimate of a response order effect that may occur in such data. Specifically, we will show that modeling response order effects as an attribute of choice alongside the substantive meaning of the ranked items in a Latent Choice Factor (LCF) model makes it possible to control for these order effects while at the same time values preferences are measured. This makes it possible to distinguish method bias effects from the content effects in which a researcher is actually interested. We will illustrate this approach using data on the endorsement of work values that were gathered by implementing a split-ballot experiment in the Longitudinal Internet Studies for the Social Sciences (LISS) panel research project. Prior to presenting the data and results of our study we review the approaches for modeling ranking data and elaborate on the benefits of the approach that we propose.

2.2 Approaches for Modeling Ranking Data

As indicated before the analysis of ranking data is not a straightforward procedure. The main statistical problems when analyzing ranking data are dependency in observation and singularity of the covariance matrix. Both problems can be regarded as sides of the same coin

that is labeled “ipsativity” in measurement. The issue of singularity reveals itself by definition when adopting an exploratory factor analysis on ranking scores per item. In this case the covariance matrix is not positive definite and cannot be estimated because a singular covariance matrix has no unique inverse. Deleting one variable, and hence row and column of the matrix, resolves this issue. One way of modeling ranking data is then by implementing exploratory factor analysis on this reduced covariance/correlation matrix. Several researchers, such as Inglehart (1977) or Kohn (1969) have adopted such an approach in the early days of researching value orientations with ranking data. However, the impact of ipsativity on the remaining associations between items in the matrix is not accounted for with this approach. The latter was solved when Jackson and Alwin (1980) introduced their ipsative common factor model. They solved the issue of linear dependency among items by allowing the errors of the items to be correlated to correct for negative correlations between the errors. These correlated errors take into account that the ranking of one item is dependent on the ranking of the other items. The Jackson and Alwin model (1980) was essentially an exploratory factor model. Chan and Bentler (1993) and Cheung (2004) further developed the model to estimate confirmatory factor models for estimating factor loadings of items that are latent in the rank-ordered ipsative data. A problem with this approach is that the rankings are being treated as if they have an underlying continuous scale, meaning that within the limits of the range of the scale they can take on any value and that the differences between two values also contains information (Allison & Christakis, 1994; Moors & Vermunt, 2007; Sacchi, 1998). However, with this method no information is used about the differences between the choices made by respondents and consequently it is impossible to give a meaning to such differences, although they are obviously crucial in a ranking task. Therefore, it seems more appropriate to view ranking items as being of an ordinal nature.

Both methods described above fail to use the full information that ranking items provide since they do not treat ranking items as such. In this paper we make use of an approach in which the actual choice process is being modeled (Croon, 1989; Vermunt & Magidson, 2005b), i.e. a Latent Class Choice model (LCC model). The model originates in the seminal work of McFadden (1986) that led to his award of a Nobel Prize. The LCC model used in this research provides an advance in McFadden's original work by allowing for different utilities to be estimated for different latent segments (Magidson, Eagle, & Vermunt, 2003; McFadden & Train, 2000), hence also controlling for measurement error (Moors & Vermunt, 2007). In addition, this model does not make assumptions regarding the measurement level of ranking items (Vermunt & Magidson, 2005a). Furthermore, an important advantage of the model is that it can be easily applied to partial rankings and that covariates can be included in the model (DeCarlo & Luthar, 2000; Moors & Vermunt, 2007). Finally, – and important for the purpose of our study – this model makes it possible to control for the response order effect by including this as an attribute of the choices respondents have to make. In other words, information about the location of the item in the choice set is operationally defined as an attribute of the ranking item.

LCC modeling has been rarely implemented in social science research applications so far. The few examples we came across (DeCarlo & Luthar, 2000; Moors & Vermunt, 2007) all fit within the traditional concept of defining latent classes or clusters. This is, however, not a requirement of the model. In this research we also make use of the possibility of imposing ordinal restrictions on the latent class variable in defining an ordinal Latent Choice Factor (LCF). In the next section we will explain how this model can be applied to the analysis of ranking data.

2.3 The Latent Choice Factor Model

In this study we are interested in modeling the ranking process when the top 3 items and the least favorite one are being selected out of j items (i.e. all alternatives to choose from). Let a_1, a_2, a_3 and a_{-1} be the items selected by a respondent, with a_1 being the item first chosen, a_2 being the second choice, a_3 being the third choice and a_{-1} being the least favorite choice. Assuming that the successive choices are made independently of one another the probability of this response pattern (a_1, a_2, a_3, a_{-1}) can be seen as:

$$P(a_1, a_2, a_3, a_{-1}) = P(a_1)P(a_2|a_1)P(a_3|a_1a_2)P(a_{-1}|a_1a_2a_3), \tag{2.1}$$

which is the product of the probability of selecting item a_1 first out of the j items, times the probability of selecting a_2 out of the remaining $j-1$ items given that a_1 was first selected, times the probability of selecting a_3 out of the remaining items given that items a_1 and a_2 were already chosen, times the probability that item a_{-1} is being chosen as the least favorite out of the remaining items given that items a_1, a_2 and a_3 were already chosen. The next step for deriving this probability is to follow the random utility model. According to this model we are able to estimate a utility μ_{a_j} for each item, where a higher value of the utility for one item compared to another means that this item has a higher ranking (Allison & Christakis, 1994). The response pattern shown above can then be determined by a logit model:

$$P(a_1, a_2, a_3, a_{-1}) = \frac{\exp(\mu_{a_1})}{\sum_T \exp(\mu_{a_t})} \times \frac{\exp(\mu_{a_2})}{\sum_S \exp(\mu_{a_s})} \times \frac{\exp(\mu_{a_3})}{\sum_R \exp(\mu_{a_r})} \times \frac{\exp(-1 * \mu_{a_{-1}})}{\sum_Q \exp(-1 * \mu_{a_q})}. \tag{2.2}$$

In this case the value μ_{a_j} can be seen as the degree to which a respondent prefers item a_j over all other items, where T is the original set of j items, S is the remaining set of items minus the one item chosen first, R is the item set minus the items ranked first and second and Q is the item set minus the top three items. The choice of the least favorite item is negatively related to

the utility of the item, in contrast to the three most favorite items which are positively related with utility of the item. To make this possible, scale weights are created with the value of +1 when the item is one of the most favorite rankings and -1 when the item is seen as the least favorite one. By taking the exponent of μ_{a_j} we can determine what the odds is that an item is being chosen out of a set of possible items.

In equation 2.2 only the pattern of choice preferences is being modeled. In the current study, however, we assume that there is a latent variable that influences these choice preferences. To model this we allow the utilities μ_{a_j} to differ over the levels of the factor(s) (i.e. the categories of the latent variable). So, each factor has its own value for each of the utilities of one item over the other items. Let us assume we have one underlying latent variable or factor, called θ_1, which is of an ordinal measurement level. The probability of showing the response pattern of selecting a_1, a_2 and a_3 as the first, second and third choice and a_{-1} as the least favorite choice is

$$P(a_1, a_2, a_3, a_{-1}|\theta_1) = \frac{\exp(\mu_{a_1}|\theta_1)}{\sum_T \exp(\mu_{a_t}|\theta_1)} \times \frac{\exp(\mu_{a_2}|\theta_1)}{\sum_S \exp(\mu_{a_s}|\theta_1)} \times \frac{\exp(\mu_{a_3}|\theta_1)}{\sum_R \exp(\mu_{a_r}|\theta_1)} \times \frac{\exp(-1 * \mu_{a_{-1}}|\theta_1)}{\sum_Q \exp\left(-1 * \mu_{a_q}|\theta_1\right)}, \quad (2.3)$$

which shows that the choice a respondent makes now depends on the value (or level) of the latent choice factor. Equation 2.3 can also be written in a regression-like way, which is:

$$P\left(a_j|\theta_1\right) = \frac{\exp\left(\beta_{j0a_j} + \beta_{j1}a_j\theta_1\right)}{\sum_{a_j} \exp\left(\beta_{j0a_j} + \beta_{j1}a_j\theta_1\right)}. \quad (2.4)$$

In this formula β_{j1} is the category-specific loading (slope) on the factor θ_1 for item a_j and β_{j0} can be seen as the intercept for item a_j. The intercepts indicate the relative preference for each item at the lowest level of the latent choice factor, whereas the slope defines the change in

relative preference per unit change in the latent choice factor. Given that this research includes a split-ballot design in which the ordering of items was randomly assigned to different groups, the model described above can be extended to take a response order effect into account. In this research we are interested in finding out to what extent order effects reflect primacy and/or recency response bias. These types of response order effects are dependent on the placement of the item in the list of alternatives; they are the same for all respondents, meaning that it is a choice-specific trait and as such it is modeled as an attribute of the choice. A primacy effect, a recency effect, or both can be included in the model. Let z_j be the primacy and/or recency indicator and β_z the effect of this attribute of the choice. The extended version of equation 2.4 with the primacy/recency order effect then becomes:

$$P(a_j|\theta_1, z_j) = \frac{\exp(\beta_{j0a_j} + \beta_{j1}a_j\theta_1 + \beta_z z_j)}{\sum_{a_j}\exp(\beta_{j0a_j} + \beta_{j1}a_j\theta_1 + \beta_z z_j)}. \quad (2.5)$$

Finally, it is possible to include external variables or covariates in the model. In fact, the order effect specified above can be seen as a covariate. But where this order effect is a choice-specific trait, external variables like age, gender and education are individual-specific traits.

2.4 Design and Method

Use was made of the LISS panel administered by CentERdata to collect our data. This LISS panel, initiated in 2007, is a representative sample of Dutch individuals, based on a true probability sample of households drawn from the population register, who participate in monthly internet surveys. Households that did not have a computer or internet access were provided with these materials to be able to participate. Our questionnaire was implemented in the summer of 2012 in a small experiment and was send to 7425 panel members, aged

between 16 and 92, of which 5899 members responded (response rate of 79.4%). Of these respondents a smaller subsample of 2913 received the ranking questionnaire. Ten panel members out of this subsample were excluded because they did not complete the questionnaire.

For the current study a survey question from the European Values Study (EVS) 2008 was used and transformed into a partial ranking task. This question measures the importance of 17 job aspects, with most items identical to ones used in previous work values research (e.g. Knoop, 1994; Ros, Schwartz, & Surkiss, 1999). Respondents were asked which of the 17 job characteristics was most important to them, which was the second most important to them of the remaining 16 alternatives and which was the third most important to them of the remaining 15 alternatives. Last, they were asked which alternative of the list containing the remaining 14 alternatives they found was the least important to them. In each of these tasks, respondents were forced to choose only one alternative.

We implemented a split-ballot experimental design to elicit and detect a response order effect. Specifically, respondents were randomly divided into two groups which each received the items of the questionnaire in a different order. From the total 2903 respondents, 2316 received the first order (version A) and 587 received the second order (version B) of the questionnaire. The unequal split into a larger and smaller group has to do with anticipated variations in measurement in future research projects based on these data. As can be seen in Table 2.1, the order between the two versions differed by dividing the questionnaire in half (see the dotted line) and then reversing the order of the items for each half. The benefit of this approach over the approach of just reversing the items is that the placing of the items is more varied between the versions. By only reversing the items, the ones shown at the beginning and the end of the list are the same for all respondents, while it is possible that the primacy and recency response order effects can occur at the same time. We particularly chose this specific

split-ballot design since we aimed at researching primacy and recency response order effects. Our approach allows to research whether primacy and/or recency accounts for the major differences between the two rank-ordered sets used. Researchers aiming at investigating other types of rank order effects can use our approach as long as it is implemented in the split-ballot design. It is impossible to implement all possible order effects in such a design since the number of different rank orders by far exceeds the number of respondents in a study. Hence even randomly assigning respondents to one of the possible rank orders does not exclude the possibility that results are affected by the omitting certain rank orders.

Statistical analysis of the specified models was possible using the syntax module of Latent GOLD Choice 4.5 (Vermunt & Magidson, 2005b). This program is a specially developed extension of the Latent GOLD program for estimating latent variable models for choice and (partial) ranking data. To use this program, the data need to be adapted into a long-file format with one record per ranking per individual. In Appendix A details are given for constructing the data files needed to estimate the models in the previous section and in our empirical application. Goodness of fit of the models was evaluated by looking especially at the likelihood-ratio chi-squared (L^2) and the Bayesian Information Criterion (BIC) values which are the most commonly used fit indices to evaluate the fit of latent class models. L^2 decreases as more parameters are added, sometimes resulting in over fitting of models. BIC resolves this problem by introducing a penalty term for sample size and the number of parameters in the model. As such, it identifies the most parsimonious model. An additional advantage of BIC is that it can be used to compare non-nested models. In general, when comparing models the model with the lowest BIC value is preferred (Raftery, 1995).

Table 2.1 Ordering of job aspect items in two experimental conditions

	Version A		Version B
(1)	Good pay	(9)	Meeting people
(2)	Pleasant people to work with	(8)	Generous holidays
(3)	Not too much pressure	(7)	A useful job for society
(4)	Good job security	(6)	An opportunity to use initiative
(5)	Good hours	(5)	Good hours
(6)	An opportunity to use initiative	(4)	Good job security
(7)	A useful job for society	(3)	Not too much pressure
(8)	Generous holidays	(2)	Pleasant people to work with
(9)	Meeting people	(1)	Good pay
(10)	A job in which you feel you can achieve something	(17)	People treated equally at the workplace
(11)	A responsible job	(16)	Have a say in important decisions
(12)	A job that is interesting	(15)	Family friendly
(13)	A job that meets one´s abilities	(14)	Learning new skills
(14)	Learning new skills	(13)	A job that meets one's abilities
(15)	Family friendly	(12)	A job that is interesting
(16)	Have a say in important decisions	(11)	A responsible job
(17)	People treated equally at the workplace	(10)	A job in which you feel you can achieve something

2.5 Results

Traces of evidence of order effects can be seen when we inspect the average rank scores of the ranking items between the split-ballot versions. We test whether the average rank scores differ significantly using the Wilcoxon-Mann-Whitney test, which is a t-test for non-normally distributed data. In Table 2.2 these average rank scores are shown and it is visible that there exists a clear primacy effect for the first two items of version A ("good pay"; "pleasant people to work with") and the first item of version B ("meeting people"). Also the difference in average rank scores between the items "a job that meets one's abilities" and "learning new skills" is significant. The first finding suggests that including a primacy order effect into the measurement model might explain to a large extent the differences between the two conditions. We further tested whether differences in ranking originate from the four choices

Table 2.2 Mean rank scores of the ranking items

	A	*B*	*P-value*
Pay	3.28	2.79	**.000**
Pleasant people	3.38	2.99	**.000**
No pressure	2.03	2.00	.070
Job security	2.20	2.20	.589
Good hours	2.20	2.26	.095
Use initiative	2.09	2.10	.922
Useful for society	2.12	2.16	.431
Holidays	1.93	1.96	.983
Meeting people	2.29	2.67	**.000**
Achieve something	2.16	2.24	.440
Responsible job	2.08	2.14	.301
Interesting	2.37	2.33	.692
Meeting abilities	2.59	2.83	**.000**
Learn new skills	2.04	2.10	**.005**
Family friendly	1.95	1.96	.380
Have a say	1.95	1.93	.854
People equally treated	2.33	2.35	.106

Note: Values in bold indicate significantly different mean rank scores

respondents had to make. Chi-square tests indicate that the differences in the rankings of the two versions of the questionnaire are significant for the top three choices (1^{st}: $\chi^2(16) = 141.60$, $p = 0.000$; 2^{nd}: $\chi^2(16) = 71.53$, $p = 0.000$; 3^{rd}: $\chi^2(16) = 35.35$, $p = 0.004$) but not significant for the least favorite one ($\chi^2(16) = 24.12$, $p = 0.087$). Given these findings we decided to model a primacy effect in subsequent analyses in the following way: if an item was presented first or second in the list of alternatives and it was chosen as a preferred item (first, second or third most important) the primacy variable was coded "1"; in all other cases the value was set to "0".

The question that emerges from these findings is whether primacy accounts for the principal difference between the two test conditions or whether the full rank order information is necessary to control for response order effects. To investigate this further, we estimated and compared LCF models using the Latent GOLD Choice module (see Appendix A). For this we used the pooled data from the two samples that differed in presentation order of the items (order effect). We compare four models. The first model includes a content factor only, thus disregarding the split-ballot design. The choice for a one-factor model is in accordance with the literature in which an underlying distinction in terms of extrinsic versus intrinsic values is depicted (Ros et al., 1999). The extrinsic or material type of work values are indicated by the items "pay", "no pressure", "job security", "good hours", "holidays", "family friendly" and "people equally treated". The items "use initiative", "useful for society", "meeting people", "achieve something", "responsible job", "interesting", "meeting abilities", "pleasant people", "learn new skills" and "have a say" are all part of the intrinsic work value dimension. In the second model we add an order effect (split sample version) to account for the experimental design. In this model we account for any difference in order between the two versions of the questionnaire. Model three adds the primacy effect to the second model, thus indicating whether primacy adds to our understanding of relative rankings and of which we expect that it reduces the order effect. Finally we estimate a model that only includes the primacy effect alongside the content factor. If the latter model fit exceeds the fit of the previous model this indicates that the primacy effect sufficiently identifies the major difference between the two split samples and hence accounts for the order effects. The LCF model we want to estimate implies making decisions regarding the number of ordered response categories that the latent variable should contain. These categories define the levels of the latent choice factor. To decide on the number of ordered response categories we estimated models in which we varied the number of levels between two until five. Table 2.3 reports the model fit statistics of the

Table 2.3 Model fit statistics of Latent Choice Factor models with varying levels, with and without primacy and order effects

Model:	L^2	BIC(LL)	Number of parameters
2-level model			
1. Content factor only	13937.51	55198.20	33
2. Model 1 including order effect	13772.51	55160.77	49
3. Model 1 including primacy + order effect	13555.43	54951.67	50
4. Model 1 including primacy effect	**13660.70**	**54929.36**	**34**
3-level model			
5. Content factor only	13899.32	55167.98	34
6. Model 5 including order effect	13731.45	55127.68	50
7. Model 5 including primacy + order effect	13541.14	54945.35	51
8. Model 5 including primacy effect	**13639.49**	**54916.12**	**35**
4-level model			
9. Content factor only	13893.02	55169.65	35
10. Model 9 including order effect	13728.93	55133.14	51
11. Model 9 including primacy + order effect	13529.51	54941.69	52
12. Model 9 including primacy effect	**13625.38**	**54909.99**	**36**
5-level model			
13. Content factor only	13891.64	55176.24	36
14. Model 13 including order effect	13727.32	55139.50	52
15. Model 13 including primacy + order effect	13527.56	54947.72	53
16. Model 13 including primacy effect	**13624.36**	**54916.94**	**37**

aforementioned models, i.e. four models with each model having a LCF with a specific number of levels.

The lowest likelihood-ratio chi square of any series of nested models is by definition obtained when all hypothesized effects are included. In Table 2.3 this is the model in which both order and primacy effect are included. To decide whether a more parsimonious model is to be preferred the information criterion BIC (Raftery, 1995) is used. Comparison of BIC values shows that the model with only the primacy effect controlled has the best fit, taking into account parsimony of the model. Thus, primacy is the main cause of differences in relative rankings of items between the two split-ballot versions. This conclusion is consistent across all sets of models with different numbers of levels for the latent choice factor. Furthermore the BIC value of the four-level latent choice factor model with primacy effect is the lowest overall. In the remainder of our analyses we continue to use the four-level models and compare the latent choice factor model without primacy (model 9) with the model taking primacy into account (model 12).

Table 2.4 shows the parameters of the two aforementioned selected models, i.e. the content factor only model (Model A) and the parsimonious best fitting model that additionally includes the primacy effect (Model B). The intercepts β_{j0} are logit coefficients that describe the relative preference for a particular item compared to other items in the set at the lowest level of the latent choice factor (level = 0). The logits indicate relative preferences since they sum to zero, which is a property of the factor model for rankings. This also explains why most of the other values for the parameter estimates changed when we accounted for primacy instead of just a change in the first two items. The slopes β_{j1} describe the change in the logit of choosing an item when the latent variable changes one unit. The unit in this case reflects the levels since they were coded 0, 1, 2 and 3. For example, in Model A (Table 2.4) the logit

of choosing the item that emphasizes "good pay" is 0.136 when the level of the latent variable is 0, but it increases with 1.531 as the level of the latent variable goes up one level. The factor slopes β_{j1} are used to interpret the factor structure since they indicate what the effect is of moving from one level to another of the latent variable on the relative preference for the respective items. In Table 2.4 we grouped the items in descending order on this effect size of model A to facilitate interpretation. In both models it follows the intrinsic-extrinsic value distinction known from work values research, i.e. relative preferences for extrinsic work values tend to increase whereas relative preferences for intrinsic values tend to decrease. As such it means that the relative preference of extrinsic motivational aspects is increasing with increasing level of the latent choice factor. In Figure 2.1 we will illustrate how the relative composition of extrinsic versus intrinsic motivational aspects of work values is gradually changing when moving across all four levels of the latent choice factor. However, before discussing this particular issue we would like to draw the reader's attention to the findings presented in Table 2.4, which provide insight into two important issues: the change in item preference relative to its overall preference level, and the effect of including a primacy effect on intercept and slopes in the equation.

We already indicated that the slopes (β_{j1}) indicate whether the relative preference for particular items increases or decreases across levels of the LCF. Within each set of extrinsic (top five items in Table 2.4) and intrinsic (bottom six items) work values items differ in overall popularity even at the lowest 'intrinsic motivation' level of the LCF. For instance, the intercept value of the item "meeting abilities" is highly positive ($\beta_{j0} = 1.663$), indicating that it is relatively much more preferred than on average, but the item becomes less popular when moving to the next level ($\beta_{j1} = -.788$). Similarly, the intrinsic work value "have a say" is the least preferred among the intrinsic values ($\beta_{j0} = -.392$) and its popularity further decreases with increasing levels of the LCF ($\beta_{j1} = -.680$). We draw the reader's attention to

Table 2.4 Comparison of latent class factor weights of the model with and the model without primacy effect

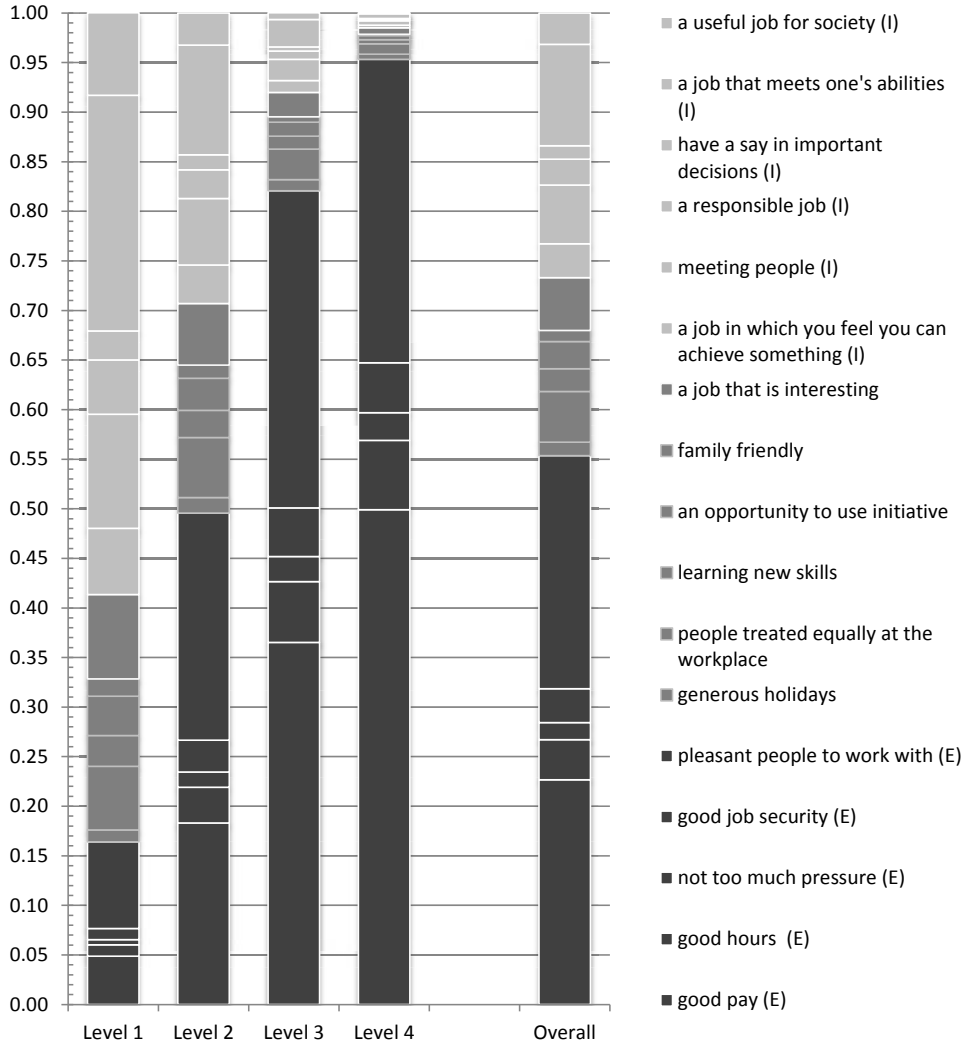
	Model A. 1-Factor model without primacy				*Model B. 1-Factor model with primacy*			
	β_{j0}	SE	β_{j1}	SE	β_{j0}	SE	β_{j1}	SE
Pay[A1]	0.136	0.260	1.531	0.193	-0.361	0.326	1.302	0.222
Pleasant people[A2]	0.834	0.184	1.028	0.116	0.220	0.249	0.945	0.163
Good hours	-0.906	0.186	1.014	0.104	-1.107	0.285	1.102	0.180
Job security	-0.982	0.177	0.946	0.103	-1.124	0.270	1.000	0.270
No pressure	-1.664	0.183	0.917	0.114	-1.912	0.291	1.065	0.187
Holidays [B2]	-1.232	0.108	0.170	0.102	-1.276	0.122	0.179	0.099
People equally treated	0.549	0.098	-0.167	0.100	0.612	0.117	-0.102	0.100
Learn new skills	-0.210	0.096	-0.184	0.086	-0.108	0.115	-0.170	0.090
Interesting	0.621	0.123	-0.231	0.127	0.898	0.165	-0.359	0.143
Use initiative	-0.009	0.098	-0.242	0.087	0.135	0.123	-0.254	0.095
Family friendly	-0.721	0.121	-0.398	0.104	-0.693	0.158	-0.313	0.122
Achieve something	0.416	0.132	-0.539	0.103	0.655	0.183	-0.589	0.131
Responsible job	0.192	0.142	-0.614	0.099	0.456	0.200	-0.679	0.135
Have a say	-0.392	0.146	-0.680	0.097	-0.173	0.213	-0.706	0.140
Meeting abilities	1.663	0.155	-0.788	0.109	1.924	0.226	-0.809	0.157
Meeting people [B1]	1.107	0.135	-0.801	0.102	0.977	0.164	-0.625	0.126
Useful for society	0.598	0.174	-0.962	0.106	0.875	0.255	-0.985	0.164
						β_z	SE	
					Primacy (ref.=0)	0.819	0.048	

Notes: Model estimates (β_{j0}, β_{j1} and β_z) consistent with equation (5); Items are grouped in descending values of β_{j1} in model A; [A1] Item offered first in condition A; [A2] Item offered second in condition A; [B1] Item offered first in condition B; [B2] Item offered second in condition B

these findings to illustrate that the lowest level of the LCF needs not to have all items of a particular type of work values, e.g. extrinsic work values, as the least preferred at the lowest level. The best way of interpreting the meaning of the LCF is to examine the change in preference structure as indicated with the slopes. Examining changes in conditional probabilities across all levels of the LCF, as we will present in Figure 2.1, further increases our understanding of the findings.

Comparison of models A and B provides insight into the effect of adding a primacy effect to the model of work values preferences. In interpreting the effects it is important to keep in mind that items "pay" and "pleasant people" defined the top two in the list in condition A and "meeting people" and "holidays" in condition B. The primacy effect of β_z = 0.819 itself indicates the increase in ranking when items are positioned first or second in the list compared to being presented further in the list irrespective of the content of the items. More important is the impact this primacy response effect has on estimated intercepts and slopes in predicting relative preferences of items. The general finding is that controlling for primacy in model B decreases both intercepts of three of the four items that were positioned first or second in one of both versions of the questionnaire. The one exception is the item "holidays" of which the overall preference is very low anyway and for which the preference level does not change significantly across levels of the LCF. As such it does not contribute to the distinction between intrinsic versus extrinsic work values anyway. Most significant is the reduced preference of "pay" and "pleasant people" at the first level of the LCF (β_{j0}) to the extent that the relative preference of the extrinsic work values is outweighed by the relative preference of intrinsic work values at the lowest level of LCF. This observation was less clearly observed in model A that does not take into account primacy effects. To further illustrate this finding we calculated the conditional probabilities of the work values per factor

Figure 2.1 Estimated conditional probabilities (model 4B) associated with the intrinsic (I) versus extrinsic (E) work values items per level of the latent choice factor and overall



Note: items order in descending order of β_{j1} (Table 2.4 model B)

level and presented them in Figure 2.1 in comparison to the overall conditional probabilities of the items. These probabilities sum to 1 within each column.

To facilitate interpretation of Figure 2.1 we ordered the items in ascending order regarding the impact of the LCF on the slopes. Items that contribute most to the intrinsic side of the LCF are presented on top whereas extrinsic items are presented at the bottom of the figure. The lighter shades can be recognized as intrinsic work values items while the darker shades are the extrinsic work values items. Based on the shades for each level, it can be seen that the first level of the latent choice factor reveals relative higher preferences of intrinsic work values relative to extrinsic work values. This gradually changes into levels that have increasingly higher probabilities of extrinsic work values being preferred. Only within the first level of the LCF we observe that intrinsic work values outweigh extrinsic values with the sum of the conditional probabilities of the six intrinsic work values almost equal to 0.60 and equivalent sum of the five extrinsic items is equal to 0.16. At the highest level of the LCF intrinsic values are hardly preferred (less than 0.02 in sum) whereas the conditional probability of "pay" alone already equals 0.50 and the sum of the five extrinsic work values is equal to 0.95. The contrast between the two extreme levels of the LCF is high, but to put findings into perspective we should take the relative distribution of respondents across the four levels of the LCF into account. The first 'intrinsically motivated' level includes 13.3% of all respondents whereas the last 'extrinsically motivated' level only 7.1%. Levels 2 and 3 respectively represent 58.2% and 21.4% of all respondents.

2.6 Conclusion and Discussion

In this study, we have investigated whether it matters in which order one presents the response alternatives in a ranking task for measuring a latent construct; and if it matters, how

large this response order effect is and how the preference structure in ranking data changes once the response order effect is controlled. To answer these questions, we implemented a split-ballot experiment in a Dutch nationally representative survey on work values. An initial inspection of the relative rankings of choice alternatives to measure these work values indicated evidence for the existence of a primacy effect in these data. Based on this finding, we then proceeded to statistically model the latent structure of the work values with the LCF model. This modeling approach provides a straightforward way to research order effects as well as to estimate to what extent order effects reflect primacy response bias. An extrinsic versus intrinsic work value distinction in the factor weights of these estimated models was obtained that is similar to latent structures of work values found in previous research (Knoop, 1994; Ros et al., 1999). This analysis yielded a statistically significant estimate of a primacy response effect. Furthermore, we found that differences in order – capturing the experimental design of our study – were sufficiently identified by this primacy response order effect. Finally, the results indicated that the rank order of alternatives changed, after controlling for the primacy effect.

It can be concluded from our findings that a ranking measurement of work values will most likely be biased as a result of the order in which response alternatives are being presented to respondents. However, in that case the validity-threatening impact of a response order effect can be sufficiently controlled by using an adequate split-ballot design and a statistical modeling approach. This brings us to the practical implication of this finding to survey research practice: To adequately measure the preference structure in the data and to get an estimation of the response order effect, the minimum requirement is to implement two versions of the same ranking assignment with differential ordering of choice alternatives and randomly assign respondents to either condition. Then, applying the statistical approach presented in this contribution adequately reduces the negative effect of response order effects

on the measurement of interest. Considering the costs of survey implementation, developing a limited split-ballot design with differential item ordering as in the current study is more cost-effective than developing a much more elaborate design in which the presentation order of the items is fully randomized. One might assume that full randomization equally reduces the negative effect of response order effects. Undoubtedly it will reduce such bias but taking into account that a sample size is much smaller than the total number of unique orders of 17 choice alternatives (= 17! = 355 687 428 096 000) there may still be some hidden bias in the data. Modeling bias thus becomes an attractive alternative to full randomization. The subsequent statistical analysis of ranking data with the help of the LCF model has the benefit that the researcher acquires an empirical estimate of the size of the response order effect; also, it makes use of the full ranking information in the data by indicating the relative change in preference of an item when moving from one level of the latent choice factor to the next, as well as to keep track of the importance of that item within the given set. This makes the latent choice approach highly informative about the latent structure of measurements that are based on ranking assignments.

The LCF model described in this study is very flexible and can also be adjusted in several ways. First, the model allows for any type of response order effect to be included. The requirement is that the hypothesized response order effect should be implemented in the split-ballot design of the study. In fact our design allowed to research recency effects alongside primacy effects but we found no evidence in of recency effects in our data. For researchers who suspect that particular response order effects may exist, our approach is a very useful tool since it also allows for getting an estimation of the size of the response order effect. Second, the model can be applied to all possible ranking tasks (see also Vermunt & Magidson, 2005b), e.g. full ranking, best-worst ranking, etc. Third, the model allows for the inclusion of external

variables (covariates) as well, resulting in a SEM-like model. All these aspects make the LCF model very flexible in its use.

Although it appears that with the inclusion of only two versions of the same ranking assignment with differential item ordering we succeeded in adequately measuring the preferences structure in work values, one may wonder whether the inclusion of more different orders would automatically rule out the response order effect. We would still expect that a response order effect could be present in each of the versions administered, especially in the case where the more popular items are shown at the beginning or the end of the list of items. Our method only allows testing for hypothesized order effects as implemented in the split-ballot design. The method allows to estimate whether the differences between the rank orders implemented in the design can be attributed to the hypothesized response order effects. Our method does not eliminate other sources of order effects that might have an impact on the results. Full randomization is often used to eliminate order effects by design. Although full randomization has the benefit of bigger heterogeneity in the total sample it also does not include all possible order effects that is equal to 17! in this study. Further research would be needed to examine whether randomization of the item order actually rules out the response order effect.

In social science research little methodological guidance is found about how to properly analyze (partial) ranking data. Hopefully the current approach gives researchers an example of how this kind of data can be analyzed in a straightforward and more informative manner by using a modeling approach that has been specifically developed to deal with ranking data. Of course the use of a ranking approach should ultimately depend on theoretical foundations about the construct one is interested in, but in values research it was already suggested that a ranking task is the most discriminating procedure which retrieves the most information about people's personal values (DeCarlo & Luthar, 2000; Maio et al., 1996;

Ovadia, 2004).With respect to the investigation of response order effects in ranking items, we advise that researchers at the designing phase of the questionnaire think carefully which order effects may be influencing the results and that they develop a split-ballot design in which these order effects can be investigated by using the approach described in the current study.

APPENDIX A: Example data layout and Latent GOLD Choice 4.5 syntax for Latent Choice Factor models with a response order effect

For the current application of the latent class choice model in Latent GOLD three files were needed. We will show an example of each of these files:

1. **Response file**

Respondent ID	Choice	Scale weight	Version_AB	Rank123	Gender	Age
1	12	1	A	1	Male	20
1	3	1	A	1	Male	20
1	13	1	A	1	Male	20
1	9	-1	A	0	Male	20
2	11	1	B	1	Female	55
2	1	1	B	1	Female	55
2	14	1	B	1	Female	55
2	8	-1	B	0	Female	55

This file contains a row for each choice a respondent makes. The scale weight (sweight) shows whether the item is one of the most favorite items (+1) or the least favorite one (-1). Latent GOLD will recognize the top three choices by the order in which they are placed in the table (so item 12 is the first choice for respondent 1, item 3 the second choice and item 13 the third choice). The version_AB variable makes it possible to compare the two orders with each other. The rank123 variable indicates which items were the top three choices, which we needed to be able to account for the primacy effect of these items. Finally, covariates can be included in the model, like for example gender and age.

2. Alternatives file

Alternative ID	Alternative number	Primacy
1	1	1
2	2	1
3	3	0
…	…	…
15	15	0
16	16	0
17	17	0
18	1	0
19	2	0
20	8	1
21	9	1

The alternatives file makes it possible to specify for which items the primacy effect should hold (in this case the first and second). Alternative IDs 18, 19, 20 and 21 are specified for version B of the questionnaire because this version differs in order of the items and so different items are shown first. The dots between alternatives 3 and 15 indicate that all items lying in between have the same primacy value of zero.

3. Sets file

Version	Alt#1	Alt#2	Alt#3	Alt#...	Alt#7	Alt#8	Alt#9	Alt#10	Alt#...	Alt#17
A	1	2	3	…	7	8	9	10	…	17
B	18	19	3	…	7	20	21	10	…	17

The sets file makes clear that items 18 and 19 in version B of the questionnaire were similar to items 1 and 2 in version A. The same accounts for items 8 and 9 in version A and items 20 and 21 in version B. Because this information is only needed for the items possibly suffering from a primacy effect, all other items are coded identically for the two versions.

Latent GOLD's syntax module was used to estimate the one-latent class factor models with and without the inclusion of the primacy effect. The variables and equations sections of this syntax are as follows:

```
variables
caseidrespondent_ID;
repscale sweight;
choicesetid version_AB ;
dependent Choice ranking;
independent Rank123;
attribute _Constants_, Primacy ;
latent
DFactor1 ordinal 4 scores= (0 1 2 3);
equations
DFactor1<- 1 ;
  Choice <- _Constants_ + _Constants_ DFactor1 + Primacy Rank123 ;
```

In the variables section, one has to provide relevant information about the dependent (ranking of in this case four choices), independent (the variable indicating the top three choices), attribute (a constant/intercept and the variable showing which the items were placed first or second in the list of items) and latent (an ordinal factor with 4 levels) variables used in the analysis. Most of these variables have to be defined in the response file, with the exception of "Primacy" which is a variable in the alternatives file (see the description of the alternatives file in this Appendix). Also some identification variables are needed like the respondent ID, the scale weight (sweight) indicating whether the item is one of the most favorite items or the least favorite one, and the choiceset ID making it possible to distinguish the two questionnaires from each other.

The first equation defines the logistic regression model for the latent variable, which only contains the intercept ("1"). The second equation defines the regression model for the choice variable. In this case "_Constants_" refers to the intercept or β_{j0} in equations 2.4 and 2.5, "_Constants_ DFactor1" refers to the category-specific loading of the latent factor or β_{j1}

in equations 2.4 and 2.5 and “Primacy Rank123” refers to effect of the primacy effect on the choice or β_z in equation 2.5.

The syntax shown above is for estimating the model with a primacy effect. The model without this effect can be simply estimated by excluding “Primacy Rank123” from the equation. Also, the addition of covariates to this model can be easily done by specifying them as independent variables (include measurement level when not numeric) and including them in the equations.

CHAPTER 3

Comparison of Ratings and Rankings for Measuring Work Values Preferences: A Latent Class Segmentation Approach*

Abstract

A continuing discussion in sociological survey research concerns whether social values should be measured using either a rating or rather a ranking response format. The form-resistant hypothesis states that differences in the latent preference structure revealed by both approaches should be small when typical features of each format are considered. Previous research, however, has shown mixed results. We suggest that adopting a latent class segmentation approach helps to explain these mixed results: It may identify segments in the population with a similar item preference structure – regardless of whether rankings or ratings are used –, as well as segments that are linked to one format only. We apply our approach to a Dutch nationally representative survey on work values with a split-ballot design. In both the rating and ranking assignment we find two segments reflecting the intrinsic and extrinsic work values preference structure. At the same time other preference structures defined segments that differed between modes. In line with the form-resistant hypothesis the results suggest the same latent preference structure has guided particular segments in a population to respond similarly to rating and ranking questions.

* This chapter is submitted for publication in Journal of Official Statistics.

3.1 Introduction

The majority of studies on attitudes or social values use rating questions in which respondents are either asked to indicate how important particular issues are, how much they agree or disagree with statements, how satisfied they are with situations, or how much something applies to them. Scales are then developed in which a set of items is assumed to measure the same underlying attitude or value, for example reflecting levels of approval of certain views or satisfaction with certain states. Fewer studies make use of ranking questions in which respondents do not evaluate each statement separately but jointly. Ranking involves asking respondent to indicate their preferences by choosing a statement that is most important to them, second most important to them, and so on, in a set of statements. In particular research situations both question formats can be applied to the same concept one intends to measure. In this study, for instance, we analyze work values as an exemplary case. Issues relevant for measuring work values such as "good pay", "having a say in important decisions" or "job security" can be asked as a list of items that should be rated in terms of how important each of them is; or they can be presented in a full list in which respondents need to indicate what they think is most important, second most important, and so on. Previous research that compared both measurement methods provides mixed results about the similarity of the preference structure that both methods reveal. Krosnick and Alwin (1987, 1988) were among the first to compare both methods formally, arguing that both formats should not lead to fundamentally different results if particular features of each method are taken into account. They labeled this idea the "form-resistant correlation hypothesis" which holds that observed correlations among values should remain invariant across different measurement methods. The results of their studies (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988) showed that the two measurement methods were less different than usually assumed but still too different to infer form-resistance. Other researchers comparing both methods even more boldly

concluded that these two measurement methods yield different results and that they measure fundamentally different things (Maio, Roese, Seligman, & Katz, 1996; McCarty & Shrum, 2000; Ovadia, 2004). Should we then abandon the idea of form-resistance when comparing results from rankings and rating questions and accept they are fundamentally different? Not necessarily, as we will show in this study. The issue is that holding on to a strict definition of form-resistance leads to a trivial expectation: The answer is known already since each method has unique features and therefore they can never produce identical results. One unique feature of ratings, for instance, is that respondents can rate each item equally whereas in the ranking format respondents are forced to make choices between these 'equally valuable' items. Ranking data, on the other hand, say little about how important issues are since two respondents with the same rank order of items might disagree on how important the issues are overall. In this study we argue that a latent class segmentation approach that allows transforming ratings into relative preferences results into even more similarities between the two question formats than suggested by Alwin and Krosnick (1985). The segmentation approach identifies latent classes of respondents that reveal a similar preference structure in the set of items. The method also allows taking overall agreement tendencies into account for ratings and response order effects for rankings. We will demonstrate that the method permits identifying different segments (i.e. subgroups of respondents) in the population: Segments that reveal a similar preference structure in work values – irrespective of whether rating or ranking data are analyzed – or other segments that are more specific to either ranking or rating. With allowing for the latter we test a less strict form-resistant hypothesis.

We illustrate the usefulness of testing the form-resistant hypothesis using this latent class segmentation approach with data on work values gathered in a split-ballot experiment in the Longitudinal Internet Studies for the Social Sciences (LISS) panel. In what follows we first briefly review the literature on the measurement of work values and link it to the

discussion on values reflecting either relative preference or absolute agreement. After that the choice of method and design is elaborated. We explain why a latent class segmentation approach might reveal more similarities between ratings and rankings than is recognized so far. Next, we report the empirical findings of our analyses. Finally, we discuss the results and the value of our method within the debate on using ratings or rankings.

3.2 Measurement of Work Values

The choice of measurement method should always depend on the definition of the latent construct one intends to measure. A field in which there is an ongoing discussion about which measurement method is preferable is the field of values research. Within this field, some researchers follow Kluckhohn's view that values are "conceptions of the desirable" (Parsons & Shils, 1962, p. p. 405). This means that the interest lies in how desirable each value is in its own right without directly having to make relative comparisons between values, indicating that the rating method is the best method to use. Others, however, follow Rokeach's view that "a value is an enduring belief that a specific mode of conduct or end-state of existence is personally preferable to an opposite or converse mode" (Rokeach, 1973, p. p. 5) calling for a ranking procedure. Therefore in this research area both rating and ranking procedures are used to measure social values. Whenever items reflect distinct views on the concept being measured either choice – rating or ranking – can be used. Work values, the exemplary social values case studied in this contribution, fits within this perspective.

Most studies on work values distinguish between intrinsic (task-related) and extrinsic (job benefits unrelated to the job task itself) aspects of work, although the exact labeling can be different (Elizur, 1984; Elizur, Borg, Hunt, & Beck, 1991; Furnham, Petrides, Tsaousis, Pappas, & Garrod, 2005; Super, 1962). Some studies identify a third type of work values that

reflect social or relational aspects of work (Elizur, 1984; Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Knoop, 1994; Ros, Schwartz, & Surkiss, 1999; Super, 1962). Not all studies recognize the latter type. Also, disagreement exists on the number and content of any extra factors added to the two-folded classification in intrinsic versus extrinsic work values.

To measure work values most researchers used rating items (Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Ros et al., 1999; Wollack, Goodale, Wijting, & Smith, 1971); only a few early studies used ranking items (Elizur, 1984; Super, 1962). Most notably, Ravlin and Meglino (1987) argued that work values may be so highly socially desirable that it may be necessary to use rankings rather than ratings regardless of how difficult the ranking task might be. The finding that ranking of work values is much less used than ratings is in line with nowadays' common practice in the social sciences: Rating questions are preferred mainly because they are more easily administered, answered by respondents and analyzed. Since work plays a central role in people's lives, we believe that responses to rating questions about work values will show a tendency that respondents value all work values as important. This agreement tendency confounds the true meaning given to work values and therefore it is more informative to look which values are being preferred over other values. Here we propose a method that allows modeling the preference order of work values while controlling for agreement tendency in the set of rating questions. Rather than abandoning ratings altogether – as suggested by Ravlin and Meglino (1987) – we use current methodology to rank the ratings (Magidson & Vermunt, 2006; Moors, 2010).

When comparing measurement methods it is also important to know whether the relationship of covariates with the measurement differs across methods. Even if measures slightly diverge, then similarity in how these measures relate to covariates would lead to the same sociological conclusions. This argument can be unequivocally linked to the form-

resistant hypothesis, which leads us to expect similarity in covariate effects on the measured values irrespective of the format used. Although this reasoning sounds logically, empirical reality tempers our enthusiasm by revealing mixed findings on covariate effects, even when the same or similar response format to measure work values is used. With respect to age, for instance, de Witte, Halman and Gelissen (2004) found that older respondents are more drawn towards intrinsic work values. However, in earlier research Halman (1996) found that younger respondents prefer both intrinsic and extrinsic values more strongly than older respondents; at the same time he argued that the impact of age should not be exaggerated. It remains unclear whether the findings on age should be interpreted as an age effect or rather as a cohort effect. There are also mixed findings with respect to gender. Some studies did not find a gender effect (De Witte et al., 2004; Furnham et al., 2005), while other studies found that women prefer social work values more than men (Duffy & Sedlacek, 2007; Elizur, 1994; Kashefi, 2011) and men prefer extrinsic work values more than women (Duffy & Sedlacek, 2007). Results are more consistent when socio-economic indicators are involved. A positive relationship of education with intrinsic work values has been documented (De Witte et al., 2004; Halman, 1996; Kashefi, 2011). De Witte, Halman and Gelissen (2004) also found a positive relation of income with preference for intrinsic work values. We found no research contradicting the latter findings.

3.3 Relative Preferences versus Absolute Level of Agreement

When testing the form-resistant hypothesis Alwin and Krosnick (1985) had to deal with the ipsative nature of the ranking data to use the traditional factor analytic approach for measuring work values. Ipsativity of ranking data refers to that ranking of one item by design has an impact on the ranking of other items in the given set. As a result the given set of

responses will sum to the same total for all respondents. To illustrate these features: Assume a question in which respondents need to fully rank three items. If a first choice is made the remaining two items can only be ranked second and third. With two items ranked, the ranking of the third item is fixed. Assigning rank scores "1", "2", and "3" to the respective choices implies that the sum of all choices equals "6" for all respondents. Alwin and Krosnick (1985) apply the Jackson & Alwin ipsative common factor model (1980) which imposes a set of constraints to correct for ipsativity. In this way ranking data are made equivalent to rating data. However, as a consequence they do not model preference structures directly. A first difference between our approach and Alwin and Krosnick's approach is that we directly model the preference structure of ranking data and apply a model-based transformation of rating items as well with which the relative preference structure within a set of rating questions can also be researched. The latter approach will overcome the problem also often found in consumer research, namely that overall liking tends to dominate the results of rating items, instead of measuring preference differences between the given items (Magidson & Vermunt, 2006). For work values, it could be expected that respondents will show a tendency of rating almost all items as equally important, thus expressing a general attitude toward work rather than a specific orientation. For such items it is more informative to look at the relative differences in importance instead of gauging the absolute level of importance.

A second difference between this study and previous research is that we use a latent class approach to identify discrete latent segments (or classes) in the population rather than to define latent factors. We argue that our approach helps to explain the mixed findings on the comparability of results from rating versus ranking data. The model we adopt allows identifying latent segments (i.e. subgroups of respondents) in a population that are similar between both modes as well as segments that are diverse or even unique to one question format. Thus we perform a less restrictive test of the form-resistant hypothesis. Details on

how latent class segmentation works with ranking and rating data are provided after presenting the study design.

3.4 Controlling for Response Bias in Rating and Ranking

Response bias typical to each question format might be a cause for differences in outcomes between ranking and rating measures. Krosnick and Alwin (1988) provided evidence that part of the rating-ranking discrepancy could be explained by taking the level of non-differentiation in rating items into account. In ranking items the ordering of the response alternatives may influence the results. This response order effect can become visible in the beginning or at the end of the item list (Krosnick, 2000). In this study, both types of response bias – non-differentiation and response order effects – will be examined with its corresponding measurement method in the data. Since rating data are modeled to reflect relative preferences of particular items over other items in the set, the latent class segmentation model might reveal a segment that shows non-differentiation if present in the data. For a ranking assignment we can model response order effects because we implemented a split-ballot design with different ordering of items per group. How this works will become clear when we elaborate on the statistical model in more detail later in this contribution.

3.5 Design and Data

Our data were collected by making use of the LISS internet panel administered by CentERdata. This panel is based on a true probability sample of households drawn from the population register. Households that did not have the materials to be able to participate in the web research (like a computer or internet access) were provided with these materials.

Participants of the LISS panel receive monthly internet surveys. The questionnaire used in this study was implemented in a small experiment in the summer of 2012. Out of the 7425 panel members who received the questionnaire, 5899 questionnaires were filled in (response rate of 79.4%). This sample was a priori randomly split into two subsamples with one subsample of 2913 respondents who received the ranking questionnaire and the other subsample of 2986 respondents who received the rating questionnaire. Only respectively ten and eighteen respondents did not fully complete the questionnaires and were excluded from the analyses.

For measuring work values we used a survey question which measures the importance of 17 job aspects as implemented in the European Values Study (EVS) 2008. We transformed this question into a partial ranking task and a rating task. We report the items used in this question in the upper part of Table 3.1; they are similar to ones used in previous work values research (e.g., Elizur et al., 1991; Furnham et al., 2005; Knoop, 1994; Ros et al., 1999). The question format for the ranking and rating tasks and the scale used for the rating task (endpoint labeling only) are shown in the lower part of Table 3.1. A split-ballot design was implemented for both ranking and rating procedure in which respondents were randomly assigned to one of two conditions. Conditions were further defined by different ordering of items in the questionnaire. For the ranking questionnaire this meant that 2316 respondents received a version with the original ordering of items (version A) and 587 that received a version with the changed ordering (version B). Of all respondents that received the rating questionnaire, 2391 filled in the version with the original ordering of rating items (version A) and 578 filled in the version with the adjusted ordering (version B). We deliberately chose for an unequal split between these versions in view of future research projects based on the current dataset.

Table 3.1 Questionnaire design

Ordering of job aspect items in two experimental conditions		
	Version A	**Version B**
(1)	Good pay	(9)
(2)	Pleasant people to work with	(8)
(3)	Not too much pressure	(7)
(4)	Good job security	(6)
(5)	Good hours	(5)
(6)	An opportunity to use initiative	(4)
(7)	A useful job for society	(3)
(8)	Generous holidays	(2)
(9)	Meeting people	(1)
(10)	A job in which you feel you can achieve something	(17)
(11)	A responsible job	(16)
(12)	A job that is interesting	(15)
(13)	A job that meets one´s abilities	(14)
(14)	Learning new skills	(13)
(15)	Family friendly	(12)
(16)	Have a say in important decisions	(11)
(17)	People treated equally at the workplace	(10)
Question format: ranking		
(a)	Here are some aspects of a job that people say are important. The question is which of these you personally think is the most important in a job?	
(b)	Of the remaining aspects of a job, which one do you consider next most important?	
(c)	Of the remaining aspects of a job, which one do you then consider next most important?	
(d)	And which one of the remaining aspects do you consider least important of all?	
Question format: rating		
	Here are some aspects of a job that people say are important: How important is each of these to you personally?	
	1 Very unimportant 2 3 4 5 Very important	

The split-ballot experimental design was implemented to be able to detect a response order effect in the items. This response order effect is one of the response biases that could be present in the ranking assignment. Whether similar order effects return in a rating assignment remains to be seen. We found only two studies that investigate response order effects in rating items by changing the ordering of the items. In the first study a significant primacy effect was found for only one item out of four items (Klein, Dülmer, Ohr, Quandt, & Rosar, 2004), while in the second study the response order effect was not significant (Ayidiya & McClendon, 1990). For the construction of the item ordering for version B, the original item set was divided in halves and then the items were reversed for each half. The main reason to put the middle alternatives at the start or end of the list in version B was that with simply reversing the questionnaire order the same items would be placed first or last in the list for both versions which would make it difficult to distinguish both primacy (i.e. the tendency to choose first alternatives in a list) and recency (i.e. the tendency to choose the last alternatives provided in a list) effects if both are present at the same time.

The design of each question format also contained some restrictions. In the ranking task, respondents were presented with one screen including the full list of items in which they were asked to choose the item they most strongly preferred. After providing an answer the respondent had to click on a button to go to the next screen; here the next question would become visible and the previously chosen item(s) were no longer available in the list. For the rating task, the questionnaire was constructed in such a way that respondents had to rate each of the items (all visible on one screen) one by one from top to bottom. Altering a given answer was only possible in the last response given; changing previously given answers was not possible. This was done to be consistent across both formats in that a given answer could not be altered afterwards.

3.6 The Latent Class Segmentation Approach

The approach we propose in this research deviates from what could be considered common practice: Using either exploratory or confirmatory factor analysis on a set of independently measured items. Ranking data are dependent measurements, e.g., if one knows the ranking of all but one of the items then based on this information the rank of the last item is also known. Rating data on the other hand produce independent measurements. Alwin and Krosnick (1985) made use of the Jackson & Alwin routine (1980) and demonstrated that correcting for dependency or ipsativity in ranking data allows modeling ranking data similar to how rating data are modeled. In contrast, the approach used in this study does not correct for ipsativity but actually directly models it by using the ipsative nature of the ranking data. It does so by defining latent segments in a population that differ in their relative preferences in work values (Magidson & Vermunt, 2006). Within the same framework it is possible to examine relative preferences for work values using rating data as well, as will be explained below. Furthermore, by focusing on identifying latent segments concerning work values preferences we deviate from the usual procedure to develop continuous latent factors or dimensions. We believe that ranking and ratings by definition produce response patterns that are unique to each question format. For instance, a ranking assignment excludes non-differentiation or equal ratings across the full set of items by design. Consequently, perfect similarity in factor analytic models is virtually impossible. As we elaborate below, our approach allows identifying latent segments or classes in a population of which certain profiles surface in either type of data – rating or ranking –, while simultaneously maintaining the option to identify segments that are uniquely defined by either question format.

The statistical analysis of the ranking items is done by using a latent class choice model. Building on the work of McFadden (1986) and Croon (1989), Vermunt and Magidson

(2005b) developed a latent class choice model in which the actual choice process in ranking items is modeled (for technical details see Vermunt and Magidson (2005b) or Section 2.3 of this dissertation). In the current study, for different latent segments different utilities or preference structures are estimated (Magidson, Eagle, & Vermunt, 2003; McFadden & Train, 2000). Furthermore, previous research (Vriens, Moors, Gelissen, & Vermunt, in press) has shown that it is possible to control for response order effects in the ranking items by including these as a choice attribute in the model. This choice attribute influences the rank of the choices made by respondents independent of the content of the items being ranked; in this case the impact of being positioned first or second in a list on the ranking of the item. This means that when respondents are being influenced by a primacy effect, which has a biasing influence on the results, the choice attribute in the model controls for this effect.

An analogous model can be built for the rating items. The method involves including a random intercept in the latent class model to control for overall agreement. As such it is a model-based alternative to within-case centering of rating data (Magidson & Vermunt, 2006; Moors, 2010). Within-case centering involves subtracting the mean level of agreement for all items from the observed rating of each item. This procedure transforms the data to a continuous scale with a complicated distribution that is difficult to analyze (Magidson & Vermunt, 2006). However, the benefit of the model-based approach is that it maintains the ordinal metric of the data. The resulting effect parameters for each latent class indicate a higher (positive estimates) or lower (negative estimates) rating relative to the random intercept. With this approach it is also easy to see whether there will be a group of non-differentiators present in our dataset. These respondents will become visible in a separate latent class or segment with the item parameter estimates hardly differing from the overall importance level.

The latent class models specified above were estimated with Latent GOLD Choice 5.0 (Vermunt & Magidson, 2005b), an extension of the original Latent GOLD program for the estimation of latent variable models for choice, ranking and rating data. In latent class analysis a decision has to be made on the number of latent classes or segments that will be distinguished. One approach to arrive at this decision is compare goodness of fit statistics – usually the Bayesian Information Criterion (BIC) – and choose the best fitting model. However, it has been argued that these fit statistics keep 'improving' without necessarily identifying theoretical meaningful latent classes. For that reason we follow Hagenaars (1990) who suggested that it is safe to interpret results from analyses with fewer latent classes if adding another class does not result in important changes in the meaning of the other classes. Therefore our decision on the number of latent classes to retain is also based on the conceptual interpretation of each of these latent classes. For each of the models estimated we evaluated the interpretability of all the estimated classes. Adding more latent classes was stopped once this did not change the meaning of existing classes and the meaning of the newly added class could not be clearly interpreted. Also, associations between covariates and the latent classes were compared for each of the estimated models, while also checking whether results would not substantially alter when latent classes were altered.

The covariates or external variables used in the current study are age (in 6 categories), highest level of completed education (in 6 categories), gender and personal gross monthly income (= per 1000 Euros). We chose to include covariates which are also frequently used in previous research on work values (De Witte et al., 2004; Duffy & Sedlacek, 2007; Elizur, 1994; Furnham et al., 2005; Halman, 1996; Kashefi, 2011).

3.7 Results

Table 3.3 reports the results from the latent class segmentation approach for measuring work values with ranking and rating data. Before fitting any model we started checking for response order effects by examining the differences between the two versions with alternative ordering of items for each measurement method. For the ranking questionnaire differences were found for the first two items of version A and the first item of version B. This provided evidence for the existence of a primacy effect in the ranking questionnaire. Preliminary analyses of the ranking question in which we controlled for the primacy effect yielded a better model fit. For instance, for the selected 3-Class model (without covariates included) we found the BIC-value decreased from 54933 to 54708. For the rating questionnaire we also noted small but significant differences in average ratings between the two versions but only part of it reflected response order effects. Furthermore, a comparison of the model with and without controlling for the response order effect yielded similar results. For this reason we chose to retain the simpler model without the inclusion of response order effects in the rating model. In conclusion, while the primacy effect in the model for the ranking items explains a part of the differences between the two questionnaire versions, the response order effect does not seem to play a large role in the rating questionnaire.

As far as the number of latent classes or segments is concerned we present findings from the 3-Class ranking model and the 4-Class rating model. As stated before, the choice of models we report here depends on the statistical and the theoretical-substantive interpretation of the results for each model. Table 3.2 presents model fit statistics of models with one until six latent classes. In both datasets the largest drop in BIC is observed between the first and second model. For the ranking data defining three rather than two classes results in an additional substantial decrease after which the decrease in BIC flattens. With rating data a

Table 3.2 Model fit of latent class segmentation analyses

Ranking data			
	LL	BIC(LL)	ΔBIC
1-Class + Primacy	-27694	55523	
2-Class + Primacy	-27137	54640	-883
3-Class + Primacy	-26821	54240	-400
4-Class + Primacy	-26679	54188	-52
5-Class + Primacy	-26568	54197	9
6-Class + Primacy	-26477	54245	48
Rating data			
	LL	BIC(LL)	ΔBIC
1-Class regression + Random intercept	-56661	113499	
2-Class regression + Random intercept	-54085	108602	-4897
3-Class regression + Random intercept	-52743	106181	-2421
4-Class regression + Random intercept	-51935	104830	-1351
5-Class regression + Random intercept	-51446	104115	-715
6-Class regression + Random intercept	-51149	103786	-329

Note: Covariates included: age, education, gender and income

Δ BIC = BIC model with *k* classes - BIC model with *k-1* classes

similar pattern is observed until four classes with a reduction in decrease afterwards. The 4-Class model of the ranking data and the 5-Class model of the rating data did not alter the findings from their respective preceding models. The 4-Class model of the rating data is particular interesting from a content point of view since it clearly identifies a class of non-differentiators. Rather than imposing the same number of classes in both datasets we preferred modeling the later unique feature of rating data and therefore compare the 3-Class ranking model with the 4-Class rating model.

The latent class segmentation analysis specifies a logit regression model for the logit associated with a given ranking or relative rating for an item, conditional on the membership of a particular latent class. Since effect coding is used the β values reported in Table 3.3 sum to zero across all items. Positive β values indicate that a particular latent class prefers a particular item relatively more than average and negative β values indicate a lower than average preference. To assign meaning to the latent classes one needs to compare the β values of an item across all latent classes. This is necessary since the β value combines two pieces of information, namely: The overall popularity of an item in the given set plus the deviance from that overall popularity within the particular latent class. By comparing across latent classes we can identify the latent class with the highest preference for the particular item. To simplify interpretation we added a column (column 10) showing for which latent class the highest ranking or relative rating was observed for each item. Two latent classes return in both methods, i.e. an intrinsically and an extrinsically oriented class. Only the item "not too much pressure" is classified in the Social latent class in the ranking assignment but its β value for the Extrinsic class is only marginally smaller and not statistically significantly different from it. Therefore seven items classify as Intrinsic whereas six items are linked to Extrinsic work values in both the ranking and rating condition. The remaining four items in the set have different meanings in the ranking and rating model. The formal comparisons of β values across the Intrinsic and Extrinsic class is presented in the last column and is calculated as the β value of the Intrinsic class minus the β value of the Extrinsic class. One could consider these differences as a slope: They show the change in the logit by moving from the Extrinsic to the Intrinsic latent class. Positive values mean that the particular item is more preferred in the Intrinsic class; negative values indicate the opposite. Overall the similarity is striking: Reported values on items representing the two classes are highly similar. Perhaps a minor exception is "learning new skills" that shows a less pronounced difference in β value

Table 3.3 Latent class segmentation in work values comparing 3 class ranking model with 4 class rating model

Ranking: 3Class model*									
	Intrinsic		Extrinsic		Social			**	Intrinsic -
	β	(SE)	β	(SE)	β	(SE)			Extrinsic
A job that meets one's abilities	1.986	(0.074)	0.332	(0.096)	0.796	(0.105)		I	1.654
A responsible job	0.472	(0.090)	-0.784	(0.094)	-0.529	(0.122)		I	1.256
A job that is interesting	1.390	(0.080)	0.282	(0.108)	-0.528	(0.126)		I	1.108
A job in which you feel you can achieve something	-0.427	(0.117)	-1.481	(0.080)	-0.979	(0.115)		I	1.055
Have a say in important decisions	0.631	(0.090)	-0.396	(0.093)	-0.138	(0.116)		I	1.027
An opportunity to use intitiative	0.134	(0.093)	-0.353	(0.084)	-0.130	(0.100)		I	0.488
Learning new skills	-0.115	(0.094)	-0.411	(0.084)	-0.314	(0.099)		I	0.296
A useful job for society	0.288	(0.107)	-1.085	(0.090)	0.376	(0.108)		S	1.373
Meeting people	0.236	(0.105)	-0.312	(0.102)	1.083	(0.092)		S	0.548
Family friendly	-1.665	(0.098)	-1.194	(0.091)	-0.106	(0.112)		S	-0.471
People treated equally at the workplace	-0.225	(0.112)	0.315	(0.094)	1.287	(0.086)		S	-0.540
Generous holidays	-1.337	(0.096)	-0.832	(0.085)	-1.185	(0.099)		E	-0.505
Pleasant people to work with	0.740	(0.085)	1.967	(0.067)	1.245	(0.090)		E	-1.227
Good pay	0.989	(0.087)	2.283	(0.079)	-0.326	(0.175)		E	-1.294
Not too much pressure	-1.690	(0.099)	-0.239	(0.132)	-0.226	(0.170)		S	-1.451
Good job security	-0.656	(0.119)	0.923	(0.079)	-0.502	(0.128)		E	-1.578
Good hours	-0.753	(0.111)	0.985	(0.081)	0.178	(0.122)		E	-1.738
Class size (proportion)	0.319	(0.018)	0.427	(0.020)	0.255	(0.020)			

* controlled for primacy effects

** highest β per item (I = intrinsic; E = extrinsic; S = social)

Rating: 4Class model ***										
	Intrinsic		Extrinsic		People		Non-differentiation		****	Intrinsic -
	β	(SE)	β	(SE)	β	(SE)	β	(SE)		Extrinsic
A job that meets one's abilities	1.291	(0.068)	-0.134	(0.047)	0.544	(0.079)	0.323	(0.072)	I	1.425
A responsible job	0.320	(0.058)	-1.357	(0.050)	-0.735	(0.078)	-0.259	(0.055)	I	1.677
A job that is interesting	1.205	(0.066)	-0.338	(0.047)	0.280	(0.079)	0.224	(0.069)	I	1.543
A job in which you feel you can achieve something	-0.187	(0.051)	-1.228	(0.048)	-1.044	(0.074)	-0.211	(0.056)	I	1.042
Have a say in important decisions	0.337	(0.057)	-0.757	(0.046)	-0.216	(0.073)	0.015	(0.061)	I	1.093
An opportunity to use intitiative	0.675	(0.060)	-0.175	(0.046)	0.270	(0.076)	0.104	(0.063)	I	0.850
Learning new skills	0.425	(0.056)	-0.373	(0.046)	0.039	(0.075)	0.152	(0.065	I	0.798
A useful job for society	-0.574	(0.049)	-0.596	(0.047)	-0.415	(0.075)	-0.278	(0.053)	ND	0.023
Meeting people	-0.026	(0.058)	-0.215	(0.050)	0.313	(0.090)	-0.157	(0.057)	P	0.189
Family friendly	-1.237	(0.053)	-0.381	(0.049)	-0.819	(0.078)	-0.213	(0.057)	ND	-0.857
People treated equally at the workplace	0.838	(0.061)	1.205	(0.058)	1.209	(0.084)	0.420	(0.074)	P	-0.367
Generous holidays	-1.078	(0.051)	-0.224	(0.047)	-0.917	(0.075)	-0.275	(0.055)	E	-0.854
Pleasant people to work with	0.927	(0.065)	1.689	(0.068)	1.599	(0.089)	0.353	(0.071)	E	-0.763
Good pay	0.039	(0.054)	0.745	(0.054)	0.400	(0.084)	0.156	(0.066)	E	-0.706
Not too much pressure	-1.573	(0.059)	0.364	(0.055)	-1.031	(0.084)	-0.385	(0.056)	E	-1.937
Good job security	-0.725	(0.053)	0.941	(0.058)	0.372	(0.095)	-0.020	(0.062)	E	-1.666
Good hours	-0.657	(0.053)	0.832	(0.053)	0.151	(0.080)	0.060	(0.065)	E	-1.489
Class size (proportion)	0.267	(0.011)	0.319	(0.012)	0.288	(0.013)	0.126	(0.007)		

*** random intercept used to model overall preference

**** highest β per item (I = intrinsic; E = extrinsic; P = people; ND = non-differentiation)

in the ranking assignment. Regardless of the observed similarity there are also noteworthy differences between the ranking and rating assignment. These difference have nothing to do with the difference in preference assigned to the Intrinsic versus Extrinsic items given the latent class (row comparisons), but rather with the overall preference assigned within the given set (column comparisons). The issue of "good pay" is a clear-cut example of this. The difference in β value between the values from the Intrinsic versus Extrinsic class is respectively -1.294 and -0.706 for the ranking and rating assignment. In both situations "good pay" is associated with an Extrinsic work orientation. In the ranking assignment, however, the overall preference of "good pay" is much higher than in the rating assignment. In the ranking assignment "good pay" is even ranked third most preferred within the Intrinsic class whereas in the corresponding rating assignment its ranking is about average. Other researchers have reported differences in overall preference ranking of items when comparing ratings with ranking (Ovadia, 2004) often inferring that data are incomparable. What our research demonstrates is that although overall preference rankings between methods might differ it is possible to find similarities in the relative preference structure across segments in the population that can be clearly identified as intrinsically or extrinsically oriented.

Four items were not classified as indicating Intrinsic or Extrinsic work values. In the ranking assignment these items group into the third Social latent class. Two of these items refer directly to the content of the job ("meeting people" and "people treated equally") whereas two other items refer to external ("family friendly") or broader ("useful for society") work values. The former two items referring to the content of the job return as the anchor items of the third People latent class in the rating assignment. Also the item "pleasant people to work with" has a high ranking in this third class that is only marginally lower than its ranking on the Extrinsic latent class. The issues of "family friendly" and "useful for society" have their highest ranking in the fourth latent class (row comparison). However, given that all

β values within this class only reveal small differences from zero this pattern is consistent with what is expected when respondents do not differentiate in relative rating.

So far our argument about a less strict form-resistant hypothesis seems to hold. To add to this interpretation we also checked the effect of covariates on latent class membership in each dataset. We note that comparisons across methods must be made with care, since the measurement part of rating and ranking models yields similarities and dissimilarities. In Table 3.4 we present the results on the estimated effects of the selected covariates on latent class membership in the ranking and rating assignment. β values sum to zero per variable (column) as well as across latent classes (row). Positive values indicate that a particular category has a higher probability of being in a particular latent class; negative values again indicate the opposite. Within the Intrinsic latent class the effect of covariates reveal mainly similarities between the rating and ranking assignment. Intrinsic work values orientation is associated with the youngest age group, increases with level of education, is higher among men, and increases with income. The picture changes when comparing the effect of covariates on the Extrinsic Values latent class. There are fewer significant and less pronounced differences in the rating assignment compared with the ranking data. Extrinsic work values decrease with educational level, but less so in the rating assignment. Education is especially important in predicting Non-differentiation class membership, contrasting lower levels of education that are much more likely to be non-differentiators with higher levels of education that are least likely to be non-differentiators. Non-differentiation is also highest among the oldest age-group and least among the youngest. The age-groups of 25-34 and 35-44 contrast with the 65+ in Extrinsic work values in a similar way in both methods, with the elderly being the least Extrinsically oriented. The other age-groups differ in impact across methods. Although income decreases the probability of being classified as Extrinsic oriented

Table 3.4 Covariate effects on latent class membership

Ranking: 3Class model						
	Intrinsic		Extrinsic		Social	
Covariates	β	(SE)	β	(SE)	β	(SE)
Age						
15-24	0.610	(0.165)	0.083	(0.122)	-0.693	(0.166)
25-34	-0.245	(0.134)	0.589	(0.109)	-0.344	(0.154)
35-44	-0.322	(0.120)	0.430	(0.092)	-0.107	(0.130)
45-54	-0.098	(0.112)	0.075	(0.084)	0.024	(0.112)
55-64	-0.079	(0.112)	-0.331	(0.084)	0.410	(0.100)
65+	0.134	(0.114)	-0.846	(0.096)	0.712	(0.100)
Educational level						
Primary school	-0.722	(0.181)	0.634	(0.127)	0.088	(0.163)
Lower secondary	-0.917	(0.148)	0.683	(0.099)	0.234	(0.129)
Higher secondary	0.198	(0.129)	-0.036	(0.110)	-0.162	(0.148)
Intermediate vocation	-0.293	(0.105)	0.071	(0.084)	0.223	(0.112)
Higher vocational	0.424	(0.101)	-0.634	(0.101)	0.210	(0.122)
University	1.311	(0.193)	-0.717	(0.206)	-0.593	(0.305)
Gender						
Men	0.170	(0.059)	0.169	(0.047)	-0.339	(0.060)
Women	-0.170	(0.059)	-0.169	(0.047)	0.339	(0.060)
Income						
(per 1000 Euros)	0.361	(0.051)	-0.036	(0.043)	-0.325	(0.056)

Rating: 4Class model								
	Intrinsic		Extrinsic		People		Non-differentiation	
Covariates	β	(SE)	β	(SE)	β	(SE)	β	(SE)
Age								
15-24	0.665	(0.114)	-0.414	(0.114)	0.201	(0.112)	-0.452	(0.170)
25-34	0.023	(0.106)	0.210	(0.101)	-0.126	(0.103)	-0.107	(0.153)
35-44	-0.127	(0.092)	0.226	(0.086)	-0.124	(0.087)	0.025	(0.119)
45-54	-0.165	(0.091)	0.222	(0.081)	-0.015	(0.081)	-0.043	(0.112)
55-64	-0.174	(0.086)	0.100	(0.076)	-0.049	(0.076)	0.123	(0.098)
65+	-0.223	(0.087)	-0.344	(0.080)	0.113	(0.113)	0.454	(0.088)
Educational level								
Primary school	-0.689	(0.144)	0.297	(0.106)	-0.178	(0.113)	0.571	(0.132)
Lower secondary	-0.678	(0.098)	0.119	(0.077)	-0.020	(0.078)	0.579	(0.095)
Higher secondary	0.128	(0.101)	0.017	(0.101)	-0.057	(0.098)	-0.089	(0.144)
Intermediate vocation	-0.214	(0.089)	-0.055	(0.080)	0.099	(0.076)	0.170	(0.105)
Higher vocational	0.543	(0.080)	-0.233	(0.088)	0.123	(0.079)	-0.433	(0.123)
University	0.910	(0.127)	-0.145	(0.161)	0.033	(0.144)	-0.798	(0.249)
Gender								
Men	0.200	(0.042)	-0.120	(0.041)	-0.069	(0.040)	-0.011	(0.048)
Women	-0.200	(0.042)	0.120	(0.041)	0.069	(0.040)	0.011	(0.048)
Income								
(per 1000 Euros)	0.113	(0.021)	-0.259	(0.036)	0.029	(0.033)	0.118	(0.021)

in the rating assignment, this is not observed in the ranking task. Finally, gender differences in Extrinsic work values are opposite across assignments: Women are more extrinsic in the rating assignment and less extrinsic in the ranking assignment. Women are also much more often classified in the third Social class in the ranking assignment. This has an impact on their corresponding score on the Extrinsic class. In summary, ranking and rating produce specific latent classes which impacts on the comparison of covariates effects on latent class membership. Similarity is highest as far as the Intrinsic class is concerned and less for the Extrinsic class.

3.8 Conclusion and Discussion

In this research we set out to investigate the boundaries of when rating and ranking formats for measuring work values produce similar results. We adopted a less restrictive form-resistant hypothesis by arguing that a model developed to test similarities should also allow for inevitable differences between the two formats. Primacy – in the case of ranking data – and overall agreement and non-differentiation – in the case of rating data – are particularities of each method that are taken into account in this research.

Our approach deviated from common practice in two ways. First, rather than eliminating ipsativity of ranking data we directly model it so that we arrive at a latent measurement of work values that reflected diversity in preference ranking of response items involved. Rating data were modeled in a similar way by adopting a procedure that separates overall agreement from measuring the relative preference structure. Second, we abandoned the idea to develop latent dimensions of work values in favor of finding latent segments that group respondents with similar work values priorities. In this way we demonstrated that an intrinsic and an extrinsic work values oriented segment emerged in both types of

questionnaire format. Other segments were specific to either ratings or rankings. Given that the latent variables from both types of data produce similar as well as distinct outcomes the comparison of the effect of covariates on these latent variables was not straightforward: The impact on one latent class of a particular latent variable affects the results of the other latent classes in the measurement. Nevertheless, similarities were observed, especially in the case of the intrinsic work values segment.

Applied researchers might find satisfaction in our research since it suggests that if one is interested in measuring diversity in preferences structure one can approximate it reasonably well with rating data without the need to administer a ranking assignment alongside the rating questionnaire. It shows that rating data – on top of the common practice of trying to measure latent 'agreement' dimensions – are also suitable to measure latent 'relative preference' classifications as well. To some extent we subscribe to this idea. The only proviso we make is that it has been demonstrated that rating data might – to a variable degree – be vulnerable to satisficing behavior by respondents as expressed in, for instance, non-differentiation. In this study the group of non-differentiators was relatively small (less than 13%). The larger this group becomes the more relevant the question: How would these non-differentiators respond to a ranking assignment? Ranking forces people to make choices and think more carefully about the questions asked and therefore reduces satisficing behavior. In this research a split-ballot design for comparing ratings and rankings was used. In future research we intend to use a within-subjects design that allows researching how non-differentiators in the rating assignment react to a ranking assignment.

Our study was limited to work values orientations only. To what extent it generalizes to other values and settings remains to be explored. The methods to analyze ranking data and to estimate relative preferences in rating data with a random intercept latent class approach

are not new; existing literature has provided evidence on its performance on other questionnaires than the one used in this research. In the current study we use a quite large item set containing 17 items, but the approach is also useful for smaller item sets. The novelty of this research, next to extending the model to account for primacy effects, was showing the similarity in latent class segmentation when comparing ratings and rankings on the same set of items. Before taking this finding for granted we recommend to do some pretesting on both formats if other types of values are involved.

We hope the current approach makes it clear to researchers that a latent class segmentation approach is a very useful tool to compare different segments of respondents, even when different measurement methods were used to define these segments. As we showed, it is not necessary to have the same number of latent classes that need to be distinguished to compare the models based on ratings or rankings. Also, researchers should be aware of the usefulness of transforming ratings into relative preferences when values are being studied. Of course the choice for using this transformation should be based on an interest in the preference of one item relative to others instead of absolute level of agreement for each item. In those cases in which agreement tendencies seem to dominate the picture, the approach used in this study allows making sense of relative deviances from the dominant picture, thus mirroring ranking profiles.

CHAPTER 4

Consistency in Work Values Preferences across Questionnaire Modes: When Ratings Meet the Rankings*

Abstract

The key research question asked in this research is to what extent the respondents' answers to ranking a set of items is mirrored in the response pattern when using rating questions. For example: do respondents who prefer intrinsic over extrinsic work values in a ranking questionnaire also rate intrinsic values higher than extrinsic values when ratings are used? We adopt a modified version of the form-resistant hypothesis arguing that each questionnaire mode yields unique features that prevent it from establishing a perfect match between both modes. By adopting a unified latent class model that allows identifying latent class profiles that share a particular preference structure in both question modes we show that a large portion of respondents tend to identify similar preferences structures in work values regardless of the questionnaire mode used. At the same time the within-subjects design we use is able to answer questions regarding how non-differentiators in a rating assignment react to a ranking assignment in which non-differentiation is excluded by design. Furthermore, the consistency of the measurement model is demonstrated by adopting a measurement invariance test that shows that the latent profile approach produces robust results in a

* This chapter has been conditionally accepted for publication in Survey Research Methods.

repeated measures setting. The findings are important since – contrary to popular belief – ranking and ratings do produce results that are more similar than often thought. Our approach is highly relevant to researchers using secondary data whenever they want to identify relative preference structures in a given dataset that was asked by rating questions and hence not directly designed to reveal such preferences.

4.1 Introduction

In survey research the overwhelming mode of asking opinion questions makes use of ratings. Ratings involve respondents indicating the level of agreement, satisfaction or importance with statements. Rankings, on the other hand, are much more rarely used. Rankings imply a respondent to list his or her priorities in a given set of items rather than indicating a level of importance or agreement. In the context of values research it has been debated whether the concept of values reflects absolute evaluations of an individual's values or rather expressing a relative preference of a particular value over others. The absolute evaluation perspective follows from Kluckhohn's idea that values are "conceptions of the desirable" (Parsons & Shils, 1962, p. p. 405), while the relative preference perspective follows from Rokeach's vision that "a value is an enduring belief that a specific mode of conduct or end-state of existence is personally preferable to an opposite or converse mode" (Rokeach, 1973, p. p. 5). Measuring values from Kluckhohn's conceptualization implies using ratings, whereas proponents of Rokeach's definition of values prefer rankings. Hence, from a conceptual point of view it is suggested that ratings and rankings would fundamentally measure different things. Admittedly, Rokeach and Kluckhohn's discussion regarding the meaning of values is ancient but still highly relevant in the field of values research (Chiusole & Stefanutti, 2011;

Klein, Dülmer, Ohr, Quandt, & Rosar, 2004; McCarty & Shrum, 2000; Ovadia, 2004; Van Herk & Van de Velden, 2007).

Regardless of this theoretical view, there has been research that focused on comparing the two questionnaire modes with both proponents for the rating method (Braithwaite & Law, 1985; Maio, Roese, Seligman, & Katz, 1996; Munson & McIntyre, 1979) as well as for the ranking method (Chiusole & Stefanutti, 2011; Harzing et al., 2009; Krosnick & Alwin, 1988; Miethe, 1985; Van Herk & Van de Velden, 2007). With Jacoby (2011) we believe that, although not stated explicitly in most literature, there is a consensus that the ranking approach is better than the rating approach because the ranking approach is more in accordance with the fundamental idea of the structure of individual values. In practice, however there all sorts of difficulties, such as for instance the cognitive demand of the task or not being able to use traditional statistical techniques, that prevent the use of the ranking approach.

Most of the studies that compared the rating and ranking methods used a between-subjects split-ballot design. This means that different respondents were randomly assigned to either the rating or ranking method and that these two groups were compared with each other. Undoubtedly very valuable insights might be obtained from such an approach – as our own research presented in previous chapters has demonstrated. However, an essential question to decide whether respondents react similarly or differently to ranking versus rating assignments remains unanswered. This central question is: are there (groups of) respondents that react in a similar way to a set of items regardless whether it is asked by means of ratings or rankings? This is the central topic of our research that can most convincingly be answered by adopting an adequate within-subjects design.

There are a few previous studies that also used the within-subjects design for measuring values using the rating versus the ranking approach (Chiusole & Stefanutti, 2011;

Maio et al., 1996; Moore, 1975; Ovadia, 2004; Van Herk & Van de Velden, 2007). Compared to the design we implemented in this research we observe three disadvantages with respect to these studies. First, the rating and ranking method was applied to questions in the same questionnaire on one time-point only. Therefore, the results of the second question can be influenced by the first question because of recognition of the question. When the ranking task was shown before the rating task, Moore (1975) found that the responses to the rating question were consistently lower. Chiusole and Stefanutti (2011) found evidence for an improved discrimination in the rating task and a better reliability for both methods, when the ranking preceded the rating task compared to the opposite order. Both these studies demonstrate that responses to a question format are affected by the preceding format used on the same set of items. In this study both question formats are asked on two separate occasions and as such we avoid this crossover effect within one measurement. A second disadvantage with previous within-subjects studies is that they could only compare what happened if the same respondents got a different measurement method at both measurement occasions. None of the previous studies included the same measurement method twice. Measuring the same method twice provides more information on comparing response consistencies. How consistent do respondents answer to the same set of items when question format changes compared to when the same format is used on each occasion? Third and finally, with few exceptions (Chiusole & Stefanutti, 2011; Moore, 1975) these studies did not vary in the ordering of the rating and ranking items. The order in which items are presented in a ranking assignment can have an effect on the choices respondents make. Primacy and recency effects might bias the measurement and make comparison with ratings more difficult to establish (Becker, 1954; Campbell & Mohr, 1950; Fuchs, 2005; Klein et al., 2004; Krosnick, 1992; Krosnick & Schuman, 1988; McClendon, 1986; McClendon, 1991; Schuman & Presser, 1996; Stern, Dillman, & Smyth, 2007). Ratings on the other hand are vulnerable to non-

differentiation (Moore, 1975; Rankin & Grube, 1980). In fact, it was this issue of non-differentiation that initiated Alwin and Krosnick's research (1985) on the form-resistance hypothesis. Controlling for question format specific response biases is hence crucial to any comparison.

In this chapter we will overcome the problems of previous within-subjects studies by showing results of a within-subjects comparison of the rating and ranking method by having all four possible combinations (rank-rank, rank-rate, rate-rank, rate-rate) tested on two measurement occasions with two months in between. A novelty of our approach compared to previous research is that we use a latent class choice modeling approach that allows us to distinguish between clusters of cases that share a common preferences pattern in the ranking as well as the rating measurement. Mode specific biases such as primacy effects, in the case of the ranking assignment, and non-differentiation, in the case of the rating assignment are simultaneously modeled. The major benefit of this approach is that it allows identifying latent classes or clusters of respondents who respond similar to both the ranking and rating task while at the same time defining classes that are much less similar. Previous research adjusted the ranking data in such a way that established methods for analyzing rating data, i.e. confirmatory factor analyses and structural equation modeling, are applicable. The work of Alwin and Krosnick (1985) is exemplary for this approach. Our approach does exactly the opposite. Rating data are modeled in such a way that the analysis shows relative preferences of particular items compared to others rather than general agreement. A second difference is that we define latent classes rather than latent factors, which is a distinction similar to cluster versus dimensional approach respectively. It is exactly this combination of modeling choices with defining latent classes that reveals clear similarities in response patterns across the two measurement methods that have previously been left unidentified.

In what follows we first take a closer look at the evidence on comparing ratings with rankings from the literature. Then we present the method and our approach in an intuitive way so that even scholars who are not familiar with latent class modeling can appreciate the benefits of our approach. Having some basic notion on logit modeling should be sufficient to understand the method. After describing the setup of our data collection we elaborate on the consecutive analysis indicating how they contribute to researching similarities and differences between ranking and ratings. The two subsamples that received the same format in each method serve as a comparative basis for the subsamples that differed in task on two occasions. Furthermore the former subsamples allow to more formally test for what is known in the literature as testing for measurement invariance (Meredith, 1993). The logic of our series of analyses will become clear as we progress through presenting our approaches and results.

4.2 Rating versus Ranking

In this research we focus on the issue of work values in which respondents need to either rate or rank a list of items that they consider to be of importance in work. The usual distinction made is between intrinsic and extrinsic work values (Elizur, 1984; Elizur, Borg, Hunt, & Beck, 1991; Furnham, Petrides, Tsaousis, Pappas, & Garrod, 2005; Super, 1962) sometimes complemented with a social dimension (Elizur, 1984; Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Knoop, 1994; Ros, Schwartz, & Surkiss, 1999; Super, 1962). There are other examples of social concepts that are similar in how a distinction is made between two or more aspects (e.g. intrinsic versus extrinsic) of a global concept (work values) for instance: Inglehart's materialistic versus post-materialistic political values orientations (1977, 1990); Kohn's intrinsic versus extrinsic parental values (1977); Rotter's internal versus

external locus of control (1966) – to name some of the classics in the field. All these concepts share one thing: they refer to different – often assumed opposite – aspects of an overarching concept. It is within this context that the question regarding (dis)similarities between ratings and rankings is particular relevant.

Methodological differences between the two measurement methods play an important role in the rating-ranking controversy. The methodological benefits of the rating approach are that rating questions are easy to administer, less time-consuming, can be administered over the telephone, allow identical scoring of items and that they are easier to statistically analyze (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; McCarty & Shrum, 2000; Munson & McIntyre, 1979). A main disadvantage of the rating approach is that it is susceptible of response biases like agreement response style (ARS: tendency to always agree with every item irrespective of the item content) and non-differentiation (tendency to not really differentiate between the items irrespective of the item content) (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988). These response biases may be the consequence of satisficing behavior, which Krosnick and Alwin (1987) defined as looking for the first acceptable answer instead of going for the optimal solution. This satisficing behavior leads to a reduced quality of the data.

Contrary to ratings the disadvantages associated with using rankings have led to the under appreciation of the method. Ranking items is a more cognitive demanding task for the respondents compared to the rating approach, more time-consuming, and less easy to statistically analyze because of the ipsativity of the data (Alwin & Krosnick, 1985). Ipsativity means that the ranking of the items is dependent on one another and therefore traditional statistical techniques cannot be used (Jackson & Alwin, 1980). However, previous research on comparing ratings and rankings has shown that the ranking approach gives higher quality

and more informative data, higher test-retest and cross-sectional reliability, higher validity of the factor structure, higher discriminate validity and higher correlation validity (Krosnick, 2000; Munson & McIntyre, 1979; Reynolds & Jolly, 1980). Furthermore, since respondents are being forced to discriminate between items satisficing behavior in the form of non-differentiation is excluded by design. Maio et al. (1996) argue that forced choices may be made arbitrarily and thus produce its own satisficing bias. Note that acquiescence is not possible within the ranking design either.

A comparison of rating and ranking methods in previous research showed only limited comparability in measurement between the two methods. Both Maio et al. (1996) and McCarty and Shrum (1997) found that the results of the rating and ranking approach were similar within participants that freely differentiated using the rating approach. Krosnick and Alwin (1988) were able to solve part of the rating-ranking discrepancy by accounting for the level of non-differentiation in ratings and adjusting for ipsativity in the ranking assignment. Other researchers found that the two methods perform equally well in differentiating between extreme items, but the items that are of moderate importance behave different using the two approaches (Chiusole & Stefanutti, 2011; Van Herk & Van de Velden, 2007). What all these studies have in common is that in the end they indicate that the ranking assignment somewhat arbitrarily forces the conceptually opposite aspects – such as intrinsic versus extrinsic orientation – to be bipolar on a single dimension whereas the rating assignment defines the two aspects as separate – although often negatively related – dimensions. The contribution of our study to the literature is that we take a different look at the same issue that sheds a new light on the alleged bipolarity of two aspects of work values related items. We will show that in both ratings and rankings distinct classes of respondents can be found that clearly assign greater preference of one type of work values over the other and vice versa. We will also show that respondents do this consistently across both methods. As argued before, previous

research primarily used a between-subjects design whereas our study includes a within-subjects design as well. Different from previous research in which the ranking data are adjusted in such a way that the methods used with rating data are applicable, we adjust the rating data in such a way that the specific methods to deal with choice data can be applied in a similar way as they are used to model ranking data. The inspiration of this perspective is provided to us from consumer research (Magidson & Vermunt, 2006). Also in consumer research rating questions are the predominant method of data collection whereas typical research questions refer to what kind of brand is preferred by which segments of the population (Moors, 2010). An adequate answer to these kinds of questions is important since new products are developed toward a targeted population. One major problem with consumer data is that an overall liking tends to dominate the response pattern of respondents when ratings are used (Magidson & Vermunt, 2006). For instance, tasting different brands of cakes and rating their tastefulness is– for most consumer subjects – a pleasant experience skewing the average rating towards positive overall evaluations. The same logic applies to work values: work can be regarded as of crucial importance in the life of (most) people. From this perspective it is far more difficult to find aspects of work as not valuable as indicating that they are important. As a result scores on rating questions regarding work values tend to be skewed towards positive scale points as well. It is important to not misinterpret the meaning of this "overall liking" or "overall importance" that dominates the response pattern. We do not suggest this reflects a response bias. A tendency towards overall liking or importance is only a response bias if it is independent of the true content that is measured. This is definitely not the case with expressing an overall liking in tasting goods, nor with feeling that work is generally important and by consequence also its different aspects. There have been attempts to use within-case 'centering' as a solution to eliminate the overall response tendency in a set of rating items (Cattell, 1944; Cunningham, Cunningham, & Green, 1977). This involves

subtracting the within-case mean score in a set of items from each observed score of each item and analysing these transformed data. This approach has been criticized from a statistical point of view since it creates ipsative data (Dunlap & Cornwell, 1994; Cheung & Chan, 2002; Cheung, 2006). This means that data on different items are not observed independently of each other. More specifically, within-case centering implies that the sum of all items scores in the set is fixed to a constant equal to zero. Most statistical models require independent data though and hence are not applicable. A model that overcomes the shortcomings of within-case centering has been proposed by Magidson and Vermunt (2006) who demonstrated the usefulness of a latent class ordinal regression model with random intercept in identifying latent class segments in a population that differ in their preference structure of tasting crackers. Moors (2010) has demonstrated that this approach works well whenever a researcher's aim is to construct a latent class typology of respondents with survey data on locus of control, gender roles and civil morality. This model reflects methods developed to model sequential choice processes (Croon, 1989; Kamakura, Wedel, & Agrawal, 1994; Bockenholt, 2002; Vermunt & Magidson, 2005b). Sequential choice modelling implies the analysis of ranking data in which a first choice is made out of *K* alternatives and each consecutive choice as a choice made out of *K minus the alternative in the previous step.* This model hence requires data to be ipsative. In the following sections we elaborate on these methods used in our research. To the best of our knowledge, this research is the first attempt to compare rating and rating questions using methods developed to analyse ipsative (ranking) or ipsatized (rating) data and compare its outcome in a within-subjects design. We do not adopt this approach for the sole sake of its 'novelty' but because it does allow us to identify segments in a sample whose work values preferences is similar regardless whether ratings or rankings are used. In what follows we explain the method in some detail, describe our data and the sequence of analyses we conducted to investigate (dis)similarity in

work values preferences across measurement mode; to research the stability in such preference structure and to check the level of measurement invariance in repeated measures. The latter is a formal test on whether the measurement is truly similar in the repeated setting of the same instrument. The logic of this sequential analysis will be explained in the process of presenting the setup of each part of the research.

4.3 Latent Class Choice Modeling of Ranking and Rating Data

Lazarsfeld (1950) was the first to introduce latent class analysis as a tool to build typologies based on dichotomous observed variables and Goodman (1974) extended it for polytomous manifest variables. Current software development (e.g. Mplus, Latent Gold, lEM) has made the method accessible to applied researchers. Most readers thus probably have some intuitive understanding of the classical latent class model. Probably the best way of giving latent class analysis an intuitive meaning is by reference to cluster analysis. The principal aim of latent class analysis as well as cluster analysis is to identify classes or clusters of cases that are similar in the manifest variables. The current research makes use of the generalized framework that latent class analysis has provided to deal with choice data that are typically provided with a ranking assignment, i.e. the latent class choice model for ranking data. Furthermore, by adopting a latent class regression model with random intercept, choice preferences in a rating assignment can also be revealed. In this section we elaborate on these two models and explain how the within-subjects comparison is modeled.

4.3.1 Latent Class Choice Model for Ranking Data

The model used for the ranking data in the current study is the Latent Class Choice (LCC) model. This model is based on the work of McFadden (1986) and makes it possible to model the actual choice process (Croon, 1989; Vermunt & Magidson, 2005b). In the current study we use a partial ranking approach in which respondents needed to rank their top 3 most important work values and the least important one out of j items. Let item a_1 be the item that was chosen as the most important one, a_2 as the second most important, a_3 as the third most important and a_{-1} as the least important item selected by a respondent. Making the assumption that the successive choices are made independently of one another, the probability of this response pattern (a_1, a_2, a_3, a_{-1}) equals:

$$P(a_1, a_2, a_3, a_{-1}) = P(a_1)P(a_2|a_1)P(a_3|a_1a_2)P(a_{-1}|a_1a_2a_3). \quad (4.1)$$

This means that the probability of the response pattern is a product of the probability of selecting item a_1 out of the full list of j items, times the probability of selecting item a_2 out of $j-1$items given that item a_1 was already chosen, times the probability of selecting item a_3 out of the remaining $j-2$ items given that items a_1 and a_2 were already selected, times the probability of selecting a_{-1} as the least favorite item out of the remaining $j-3$ items given that items a_1, a_2 and a_3 were chosen already. Next, we follow the random utility model in which we are able to estimate a utility μ_{a_j} for each item. A higher utility for one item in comparison with another item means that this item has a higher ranking (Allison & Christakis, 1994). Using a logit model to determine the response pattern shown above, the equation becomes:

$$P(a_1, a_2, a_3, a_{-1}) = \frac{exp(\mu_{a_1})}{\sum_T exp(\mu_{a_t})} \times \frac{exp(\mu_{a_2})}{\sum_S exp(\mu_{a_s})} \times \frac{exp(\mu_{a_3})}{\sum_R exp(\mu_{a_r})} \times \frac{exp(-1*\mu_{a_{-1}})}{\sum_Q exp(-1*\mu_{a_q})}. \quad (4.2)$$

The value μ_{a_j} is the degree to which item a_j is being preferred over all other items by a respondent. T equals the full set of items, S is the remaining set of $j-1$ items (minus the alternative chosen first), R is the remaining set items minus the alternatives selected first and second, and Q is the item set minus the items ranked as top 3 most important items. The item that was chosen as the least favorite one (a_{-1}) is negatively related to the utility of the item. This was made possible by including scale weights which could have a value of +1 when an item was chosen as the top 3 most important versus -1 when an item was chosen as the least important one. Taking the exponent of μ_{a_j}, the odds is determined that an item is being chosen out of a set of possible alternatives.

In the current application we are interested in applying a latent class analysis in which respondents are being clustered that have a similar value preference structure. Thus, each group (latent class) of respondents has its own value for the utilities. Using the LCC model, different utilities can be estimated for different latent classes (Magidson, Eagle, & Vermunt, 2003; McFadden & Train, 2000). Equation 4.2 needs to be slightly changed to account for the differences between the latent classes and becomes:

$$P(a_1, a_2, a_3, a_{-1}|X=c) = \frac{exp(\mu_{a_1 c})}{\sum_T exp(\mu_{a_t c})} \times \frac{exp(\mu_{a_2 c})}{\sum_S exp(\mu_{a_s c})} \times \frac{exp(\mu_{a_3 c})}{\sum_R exp(\mu_{a_r c})} \times \frac{exp(-1*\mu_{a_{-1} c})}{\sum_Q exp(-1*\mu_{a_q c})}, \qquad (4.3)$$

in which X is the discrete latent variable and c is a particular latent class. The higher the value of μ_{a_j}, the higher the probability that a respondent belonging to latent class c selects alternative a_j as one of the most important items.

In the current study we will show log transformed utilities, based on the following formula:

$$\ln\mu_{a_j} = \alpha_{a_j} + \beta_{a_j c} \tag{4.4}$$

(see also: Moors & Vermunt, 2007). Effect coding is used for identification purposes, and therefore α_{a_j} can be seen as the average utility of item a_j and $\beta_{a_j c}$ as the deviation from the average utility for respondents belonging to latent class c. A positive $\beta_{a_j c}$ value means that respondents belonging to latent class c have a higher probability than average of choosing item a_j as one of the most important items. Since the $\beta_{a_j c}$ values are estimated relative to the average utility, the sum of all $\beta_{a_j c}$ values within a latent class equals zero.

Last, we are also interested in the presence of a response order effect. A response order effect is present when items that are shown as one of the first or last alternatives in the list of items have a higher probability of being chosen as one of the more important items, irrespective of the actual content. In this research we present two alternative orderings of items in a split-ballot design (more details on the design of this study are provided in Section 4.4). Since the placement of the items is the same for the respondents in each subsample, the response order effect is also forced to be the same for all respondents. This means that it is a choice-specific trait and modeled as such as an attribute of choice. Equation 4.4 needs to be extended to be able to model the response order effect and becomes:

$$\ln\mu_{a_j} = \alpha_{a_j} + \beta_{a_j c} + \beta_z z_j. \tag{4.5}$$

Let z_j be the response order effect indicator (indicates whether items are presented first or last in the list) and β_z the effect of this attribute of choice. Thus, when a response order effect is present, it can be accounted for by including an extra beta value specific for this response order effect for the items that show a response order effect.

4.3.2 Latent Class Regression Model with Random Intercept for Rating Data

The main interest in the current study is to be able to compare the results from ranking data with the results from rating data. Therefore, a model was chosen for the rating data that allows controlling for the overall agreement level and estimate latent classes that differ in their relative ratings of particular items compared to other items in the set. This model is called the latent class regression model with random intercept. The inclusion of a random intercept in this regression model makes it possible to control for the overall level of agreement or importance (Magidson & Vermunt, 2006; Moors, 2010). Specifically, with the random intercept the average agreement across rating items is modeled as it varies across respondents. The latent class regression coefficients will then indicate relative as opposed to absolute differences in importance between the items. In this research we are particularly interested in the relative preference information because this information is similar to the relative preferences obtained by using the ranking method.

As indicated before the latent class regression model with random intercept is a model-based alternative to within-case centering (Magidson & Vermunt, 2006; Moors, 2010). The benefit of using the model-based approach is that the original ordinal measurement level of the rating data is being maintained (Magidson & Vermunt, 2006) and it suits the analysis of ipsatized data.

Let Y_{ij} be the rating of respondent i of item j and let m be the discrete values of the rating Y_{ij}. Since the rating is a discrete (ordinal) response variable, an adjacent-category logit model is being defined as follows:

$$\log\left[\frac{P(Y_{ij}=m|c)}{P(Y_{ij}=m-1|c)}\right] = \alpha_{im} + \beta_{cj} = \alpha_m + \lambda F_i + \beta_{cj}. \quad (4.6)$$

This is a regression model for the logit of giving rating m instead of $m-1$ for item j conditional on belonging to latent class c. α_{im} is the intercept which is allowed to differ over individuals and is a function of the intercept's expected value (α_m and a continuous factor (F_i) which is normally distributed and has a factor loading equal to λ. β_{cj} is the effect of item j for latent class c. For the identification of the parameters effect coding is used, which leads to a sum of zero for the α_m parameters over the possible ratings and to a sum of zero for the β_{cj} parameters over all items. A positive value for β_{cj} indicates that respondents belonging to latent class c value an item as more important than average. Thus, α_{im} accounts for the overall importance/agreement level and β_{cj} gives an indication of the relative preference of an item in comparison with the average importance level.

Last, it is also possible to control for a response order effect in rating items. Again, the response order effect is modeled as an attribute of choice, which is choice-specific meaning that it has the same effect over all individuals. Extending equation 4.6 to account for a response order effect, the formula becomes:

$$\log\left[\frac{P(Y_{ij}=m|c,z)}{P(Y_{ij}=m-1|c,z)}\right] = \alpha_{im} + \beta_{cj} + \beta_z z_j = \alpha_m + \lambda F_i + \beta_{cj} + \beta_z z_j. \quad (4.7)$$

The z_j parameter indicates whether items were presented first or last in the item list and β_z is the effect of this attribute on the choice respondents make. This extra beta value is only needed when a response order effect is found to be present. A requirement is that (at least) two randomly assigned subsamples receive alternative orderings of the set of items.

4.3.3 Comparing Latent Class Assignments

In both models it is possible to assign respondents to particular classes based on their posterior membership probabilities. These class assignments then are the input for subsequent analyses in which the association between repeated measurements is investigated. We make use of a recently developed approach to adequately estimate associations in a three-step design (Vermunt, 2010; Bakk, Tekle, & Vermunt, 2013). These three steps include: (1) estimating a measurement model (as presented in section 4.2.1 and 4.2.2); then (2) calculating the posterior membership probabilities and the class assignments, which are added as new variables to the dataset; and then (3) estimating the associations between the true class memberships by taking into account the classification errors in the assigned class memberships. It has been shown that outcomes from the latter analysis may lead to severely downward-biased estimates of the associations when classification errors are not accounted for (Bolck, Croon, & Hagenaars, 2004). In the current study proportional assignment will be used as classification method, which means that respondents are treated as belonging to each of the latent classes with weights equal to the posterior membership probabilities. The adjustment method that is used is the maximum likelihood (ML) method which is the preferred option for most situations (Vermunt & Magidson, 2013).

Assume that X is the latent variable, W is the assigned class membership, c is particular latent class and y is a particular response pattern. The posterior class membership probabilities can be estimated using the following formula:

$$P(X = c|Y = y) = \frac{P(X=c)P(Y=y|X=c)}{P(Y=y)}. \quad (4.8)$$

This means that the probability of belonging to a certain latent class conditional on a respondent's response pattern can be calculated by multiplying the latent class proportions

$P(X = c)$ with the class-specific response probabilities $P(Y = y|X = c)$ and dividing the probability of having a certain response pattern $P(Y = y)$ from this multiplication. The proportional assignment to each of the classes $W = c$ implies that respondents are treated as belonging to each of the classes with weights equal to the posterior membership probabilities, $P(W = c|Y = y) = P(X = c|Y = y)$. In our study, in the three-step model, we use the proportional class assignments at multiple occasions as well as the information on the resulting classification errors to estimate the association between the true class memberships across occasions. This yields the relevant information about the consistency in results when alternative measurement methods (ratings versus rankings) are presented to the respondents. Results from the same method subsamples will serve as a comparative basis. In the next section we present our between- and within-subjects design in detail.

4.4 Design

To collect our data, we made use of the LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by CentERdata. This panel is a probability-based internet panel that participates in monthly internet surveys. The LISS panel is based on a true probability sample of households drawn from the population register in the Netherlands in 2007. Households that did not have the materials to participate, like a computer or internet access, were provided with these materials. The questionnaire used in the current study was implemented in a small experiment in the summer of 2012. Since a between- and within-subjects design was used, we had two time-points at which the questionnaire was administered. The first measurement took place in June and July and the second measurement in September and October. The time between the two measurements was at least two months for all respondents. The first questionnaire was sent to 7425 panel members, aged between 16

and 92, of which 5899 responded (response rate of 79.4%). For the second measurement the questionnaire was distributed among 5697 of these respondents. 5492 of them filled in the questionnaire (response rate of 96.4%).

Since we are comparing rating and ranking methods, the sample was a priori randomly divided into subsamples. This division led to a subsample of 1675 respondents who received the ranking questionnaire twice (subsample 1), 1035 who received first the ranking and then the rating questionnaire (subsample 2), 1104 who received first the rating then the ranking questionnaire (subsample 3), and 1678 respondents that received the rating questionnaire twice (subsample 4). One panel member for the rank-rank condition was excluded because this respondent did not completely fill in the questionnaire and one panel member for the rank-rate condition was excluded from the subsample because this respondent did not respond at the first measurement occasion.

To measure work values, a survey question from the European Values Study (EVS) 2008 was used in which respondents needed to indicate the importance of 17 job aspects. The items given to the respondents (see Table 4.1) were similar to items used in previous work values research (Elizur et al., 1991; Furnham et al., 2005; Knoop, 1994; Ros et al., 1999). The question from the EVS was transformed for the current application into a rating task and a partial ranking task. For the rating task a 5-point scale was used with only labels for the endpoints. The rating questionnaire was set up in such a way that the items had to be rated from top to bottom. Altering an answer to an item was not possible after a respondent rated the next item. In the ranking task, respondents were asked to indicate their top 3 most important items and the item that was least important to them personally out of the full list of items. Once an item was chosen as the most important one and the respondent went to the next page, which contained the next question, the chosen item was dropped out of the list of

Table 4.1 Questionnaire design

Ordering of job aspect items in two experimental conditions		
	Version A	**Version B**
(1)	Good pay	(9)
(2)	Pleasant people to work with	(8)
(3)	Not too much pressure	(7)
(4)	Good job security	(6)
(5)	Good hours	(5)
(6)	An opportunity to use initiative	(4)
(7)	A useful job for society	(3)
(8)	Generous holidays	(2)
(9)	Meeting people	(1)
(10)	A job in which you feel you can achieve something	(17)
(11)	A responsible job	(16)
(12)	A job that is interesting	(15)
(13)	A job that meets one´s abilities	(14)
(14)	Learning new skills	(13)
(15)	Family friendly	(12)
(16)	Have a say in important decisions	(11)
(17)	People treated equally at the workplace	(10)
Question format: ranking		
(a)	Here are some aspects of a job that people say are important. The question is which of these you personally think is the most important in a job?	
(b)	Of the remaining aspects of a job, which one do you consider next most important?	
(c)	Of the remaining aspects of a job, which one do you then consider next most important?	
(d)	And which one of the remaining aspects do you consider least important of all?	
Question format: rating		
	Here are some aspects of a job that people say are important: How important is each of these to you personally?	
	1 Very unimportant 2 3 4 5 Very important	

possible items to select. This means that each item could be chosen only once. Also, respondents were able to choose only one item in each of the ranking tasks. See the bottom part of Table 4.1 for the rating and ranking question formats that were used.

To be able to detect a response order effect in both ranking and rating data, different orderings of the questionnaire in a split-ballot experiment were needed. Respondents were randomly assigned to either version A or version B of the questionnaire. In version A the items were shown to the respondents in the same order as the items are ordered in Table 4.1 (see also the numbers that are placed in front of the item names). In version B of the questionnaire the item set was split in half (see the dotted line in Table 4.1) and then the order of the items was reversed for each half (see also the numbering behind each item in Table 4.1). This approach differs from previous studies, in which the items are shown in a simply reversed order. The main reason why items from the middle of the list (version A) are presented at the beginning or end of the alternative list (version B) is that it makes it possible to research primacy or recency response order effects in case they would occur at the same time. With simple reversed ordering this would not be possible.

4.5 Results

4.5.1 Preliminary Analyses

The results from our previous study (chapter 3) constitute the background of the preliminary analysis in this research. We repeated our exploratory research on whether response order effects needed to be taken into account in both the ranking and rating assignment and applied it at both time-points. The main results from this exploratory research are in line with what we previously found. We compared the mean rank/rating scores for the two versions and we

compared the model fit of a model without controlling for response order effects with a model with response order effects being controlled for. The results of these model comparisons are reported in Table 4.2. For the rating questionnaires at both time-points there are items that seem to indicate a primacy and recency effect, but more than half of the differences between version A and B cannot be explained by a response order effect. Comparing the fit of the model with and without controlling for both a primacy and recency effect, it can be seen that on time-point 1 the fit only minimally improves for the model with response order effects (BIC-value decreased from 105056 to 105047) and that on time-point 2 the BIC-value increases for the more complex model (from 95410 to 95417). A comparison of the parameter estimates for the two models, with and without response order effect controlled, showed similar results. Based on these results the choice was made to retain the simpler model (without response order effects included) for the rating data. The ranking questionnaire results showed the presence of a primacy effect. The first two items of version A and the first item of version B have a significantly higher mean rank score at both measurement occasions. Also, the results of the model fit statistics indicate a better fit of the model with a control for primacy at both time-points. At time-point 1 the BIC-value decreases from 54914 to 54689 and at time-point 2 the BIC-value changes from 53959 to 53763. These results led to the conclusion to include the primacy effect for the ranking data.

Next, we will show the results from the 3-Class ranking models and the 4-Class rating models at the two time-points. In the previous chapter we elaborated on this decision in detail. In short: the choice of models depended on the theoretical interpretation of the results for each model combined with methodological criteria. The results of these models are in accordance with work values literature in which three types of work values are being distinguished, namely intrinsic, extrinsic and social work values (Elizur, 1984; Elizur et al., 1991; Furnham et al., 2005; Kalleberg, 1977; Knoop, 1994; Ros et al., 1999; Super, 1962).

Table 4.2 Model fit of latent class segmentation analyses

Ranking data T1					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
3Class model	-27258	54914	54616	54666	50
3Class model with primacy effect	-27141	**54689**	**54385**	**54436**	51
3Class model with primacy and recency effects	-27213	54833	54528	54579	51
Ranking data T2					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
3Class model	-26781	53959	53663	53713	50
3Class model with primacy effect	-26679	**53763**	**53460**	**53511**	51
3Class model with primacy and recency effects	-26745	53894	53592	53643	51
Rating data T1					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
4Class model	-52192	105056	104553	104637	84
4Class model with primacy effect	-52183	**105045**	**104535**	**104620**	85
4Class model with primacy and recency effects	-52183	105047	104537	104622	85
Rating data T2					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
4Class model	-47373	**95410**	94914	**94998**	84
4Class model with primacy effect	-47372	95415	**94913**	94998	85
4Class model with primacy and recency effects	-47372	95417	94915	95000	85

Note: Values in bold indicate the smallest goodness of fit value

The extra latent class for the rating data consists of the non-differentiating respondents. We start by comparing the parameter estimates for similar latent classes found at each time-point and for each measurement method. Then we will show the results of investigating the association between the proportional latent class assignments to each of the latent classes at the two measurement occasions. In these analyses the measurement model is estimated separately at each occasion. As such we observe conceptual resemblance between latent structures. If one changes the measurement method from 1st to 2nd occasion this is the only way of making comparisons. When measurement methods are the same at both measurements then a more formal test of consistency is established when investigating measurement invariance. This will be the final analysis presented in this study.

4.5.2 Latent Class Comparisons

The results for the latent class analyses are shown in Table 4.3. Table 4.3.1 includes the findings on the first (T1) and second (T2) ranking measurements. In Table 4.3.2 we present findings on the two waves of rating questions. The final two columns in each of these tables contrast the effect sizes of the intrinsic versus extrinsic latent class found in both the ranking and rating data. To interpret the results the following characteristics of the results need to be kept in mind:

(a) Column wise the parameter estimates sum to zero. Positive values indicate higher than average preference for the items in the given set of items when rankings are used. When ratings are used positive values indicate a higher than average rating relative to the overall rating of items, the latter which is measured by the random intercept. Negative values mean the opposite.

(b) If one wants to assign a meaning to the different latent classes, one should compare results row wise. An item may be ranked or rated highly (positive parameter estimates) in each latent class but with different magnitude across classes. For example: in the first analysis (ranking 3Class model T1) "Meeting one's abilities" has higher than average (positive) rankings in each latent class but clearly highest in the first "intrinsic" latent class (b = 1.763) and least in the second "extrinsic" latent class (b = 0.255). "Pleasant people to work with" is also an item with positive estimates across latent classes but is highest on the second "extrinsic" latent class (b = 2.051) and lowest on the first "intrinsic" latent class (b =0.892). Although both items have higher than average preferences, "meeting one's abilities" contributes to identifying a more intrinsically oriented latent class whereas "pleasant people to work with" contributes to defining the second latent class as extrinsically oriented.

(c) To facilitate interpretation we regrouped items into three categories. The top 7 items are linked to intrinsic work values, the bottom 6 items refer to extrinsic work values and the remaining 4 items in the middle differ in meaning depending on the analysis. This regrouping is based on our empirical findings and is consistent with theoretical conceptualization.

(d) Each of the analyses includes different subsamples. There were four subsamples coinciding with the four test conditions: 'rank-rank' (subsample 1), 'rank-rate' (subsample 2), 'rate-rank' (subsample 3), and 'rate-rate' (subsample 4). Respondents were randomly assigned to one of these four test conditions.

Reading the table it can be seen that the intrinsic and extrinsic work values class is consistently observed across measurement method (rankings and ratings) and across occasions. The third latent class in the ranking assignment can be linked to social work values and is observed consistently in both first and second measurement. In the ranking assignment

Table 4.3 Comparison of the parameter estimates for the work values classes for the ranking and rating latent class models

4.3.1 Ranking 3Class model	T1			T2			T1	T2
Items	Intrinsic (I)	Extrinsic (E)	Social	Intrinsic (I)	Extrinsic (E)	Social	(I) - (E)	(I) - (E)
Meeting abilities	**1.763***	0.255*	0.908*	**2.144***	0.474*	0.869*	1.508	1.670
Responsible job	**0.333***	-0.887*	-0.573*	**0.675***	-0.959*	-0.791*	1.220	1.634
Interesting	**1.210***	0.130	-0.612*	**1.033***	0.364*	-0.264	1.080	0.669
Achieve something	**0.548***	-0.514*	-0.162	**0.814***	-0.504*	-0.594*	1.062	1.318
Have a say	**-0.557***	-1.578*	-0.955*	**-0.103**	-1.259*	-1.272*	1.021	1.156
Use initiative	**0.082**	-0.379*	-0.157	**0.233***	-0.381*	-0.131	0.461	0.614
Learn new skills	**-0.184***	-0.458*	-0.267*	**-0.126**	-0.426*	-0.239*	0.274	0.300
Useful for society	0.098	-1.137*	**0.659***	-0.100	-0.987*	**0.842***	1.235	0.887
Meeting people	0.352*	-0.275*	**1.163***	0.345*	-0.361*	**1.446***	0.627	0.706
People equally treated	0.208*	0.367*	**1.229***	0.066	0.316*	**1.041***	-0.159	-0.250
Family friendly	-1.568*	-1.178*	**0.097**	-1.638*	-0.929*	**-0.429***	-0.390	-0.709
Holidays	-1.302*	**-0.884***	-1.207*	-1.731*	**-0.803***	-1.245*	-0.418	-0.928
Pleasant people	0.892*	**2.051***	1.310*	0.874*	**1.952***	1.773*	-1.159	-1.078
Pay	1.139*	**2.279***	-1.023*	0.947*	**2.164***	-0.777*	-1.140	-1.217
Job security	-0.392*	**0.944***	-0.859*	-0.470*	**0.788***	-0.432*	-1.336	-1.258
Good hours	-0.810*	**1.112***	0.310*	-0.844*	**0.849***	0.184	-1.922	-1.693
No pressure	-1.812*	**0.152**	0.140	-2.119*	-0.298*	**0.019**	-1.964	-1.821
Class size (proportion)	0.402	0.409	0.189	0.306	0.443	0.251		
subsamples		(1) + (2)			(1) + (3)			

Notes: * parameter estimate is at least twice the standard error; Values in bold indicate for each item the highest preference value over all latent classes in each model

Table 4.3 continued

4.3.2 Rating 4Class model	T1				T2				T1	T2
Items	Intrinsic (I)	Extrinsic (E)	People	Non-diff	Intrinsic (I)	Extrinsic (E)	People	Non-diff	(I) - (E)	(I) - (E)
Meeting abilities	**1.279***	-0.136*	0.529*	0.349*	**1.459***	-0.109*	1.104*	0.134	1.415	1.568
Responsible job	**0.286***	-1.357*	-0.753*	-0.239*	**0.369***	-1.371*	-0.318*	-0.379*	1.643	1.740
Interesting	**1.168***	-0.331*	0.264*	0.252*	**1.203***	-0.428*	0.854*	-0.030	1.499	1.631
Achieve something	**0.332***	-0.771*	-0.226*	0.032	**0.470***	-0.696*	0.067	-0.196*	1.103	1.166
Have a say	-0.219*	-1.228*	-1.066*	**-0.192***	**-0.064**	-1.203*	-0.679*	-0.329*	1.009	1.139
Use initiative	**0.660***	-0.180*	0.266*	0.135*	**0.667***	-0.287*	0.438*	-0.002	0.840	0.954
Learn new skills	**0.437***	-0.397*	0.036	0.169*	**0.355***	-0.445*	0.228*	-0.011	0.834	0.800
Useful for society	-0.579*	-0.599*	-0.429*	**-0.279***	-0.497*	-0.403*	-0.657*	**-0.255***	0.020	-0.094
Meeting people	0.025	-0.237*	**0.279***	-0.158*	0.128	0.256*	**0.286***	-0.181*	0.262	-0.128
People equally treated	0.910*	1.178*	**1.218***	0.399*	0.597*	1.340*	**1.459***	0.501*	-0.268	-0.743
Family friendly	-1.144*	-0.407*	-0.817*	**-0.262***	-0.999*	-0.309*	-1.185*	**-0.076**	-0.737	-0.690
Holidays	-1.372*	**-0.205***	-0.901*	-0.284*	-1.143*	-0.233*	-1.157*	**-0.243***	-1.167	-0.910
Pleasant people	0.965*	**1.700***	1.612*	0.341*	0.435*	**1.451***	1.274*	0.723*	-0.735	-1.016
Pay	0.018	**0.775***	0.431*	0.136*	-0.117*	**0.632***	0.327*	0.240*	-0.757	-0.749
Job security	-0.725*	**0.958***	0.392*	-0.039	-0.597*	**1.046***	-0.233*	0.063	-1.683	-1.643
Good hours	-0.675*	**0.838***	0.188*	0.062	-0.535*	**0.623***	-0.213*	0.151*	-1.513	-1.158
No pressure	-2.002*	**0.399***	-1.023*	-0.423*	-1.731*	**0.136***	-1.595*	-0.111	-2.401	-1.867
Class size (proportion)	0.161	0.310	0.288	0.133	0.225	0.386	0.285	0.103		
subsamples		(3) + (4)				(2) + (4)				

Notes: * parameter estimate is at least twice the standard error; Values in bold indicate for each item the highest preference value over all latent classes in each model

the four items grouped in the middle do not define a particular latent class, although the third class seems to put greater emphasize on the "people" items. The fourth and last class identifies a class of respondents that reveal little differentiation (small parameter estimates and few significant differences) in their ratings and thus can be regarded as non-differentiators.

Having a closer look at what items have their highest relative parameter estimate across classes thus reveals the meaning that can be given to each latent class. The first latent class present in both the rating and ranking method is the intrinsic work values class. Table 4.3 shows that the items "a job that meets one's abilities", "a responsible job", "a job that is interesting", "a job in which you can achieve something", "have a say in important decisions", "an opportunity to use initiative" and "learning new skills" all have the highest probability to be preferred by the respondents belonging to this class. The item "have a say in important decisions" shows a slightly deviant result for the rating approach at time-point 1 since the parameter value associated with the fourth "non-differentiation" latent class (-0.192) is marginally higher than with the first latent class (-0.219). The difference is too small to interpret this particular result as contradiction, especially not since its value on the intrinsic class is in contrast with the very low value on the extrinsic class. Hence it is safe to conclude that the overall pattern of the ranking and rating method in identifying an intrinsic work values class is quite similar.

Although tempting, it is dangerous to compare the magnitude of the estimated parameters across occasions for two reasons. First of all results from first and second measurement refer to different samples and second it is not formally tested whether observed differences are meaningful from a statistical point of view. For instance, we observe that effect parameters tend to be (slightly) higher for most of the intrinsic items on the second

occasion. To formally test whether we should pay attention to this finding we will present evidence towards the end of this analysis section from a model testing for measurement invariance in the case that respondents received the same questionnaire twice. This test will also inform us whether we should draw attention to the fact that class sizes differ from one measurement to the other. In Table 4.3 the shift in class size might be due to differences in measurement even if they are minor as reported in Table 4.3.

The second latent class is labeled as the extrinsic work values class. The items "generous holidays", "pleasant people to work with", "good pay", "good job security", "good hours" and "not too much pressure" are all the most preferred by respondents belonging to this latent class. All of these items refer to benefits beyond the content of the job itself. One could think of the item "pleasant people to work with" as symbolizing a social aspect as well but in all four analyses its highest observed effect parameter is linked to the extrinsic class. Keep in mind that overall it is a very popular item, but most popular amongst the extrinsically motivated. Having "pleasant people to work with" is evidently linked to a job – just like any of the items listed – but it is not an inherent aspect of the job as such. That is why it is less prominent – although still relatively important – among the intrinsic oriented respondents.

Whereas the comparison across occasions within each measurement method produces very similar results, the comparison across measurement methods reveals some specific findings for some of the extrinsic items. Most noticeable is the large effect parameter observed for "good pay" in the ranking assignment, which is smaller in the rating task. It is still consistent with the theoretical expectation that this would be part of the extrinsic qualification of work values, but the difference is pronounced. "Good pay" is clearly ranked highest in the ranking data but not in the rating data. Keep in mind that this result is after controlling for primacy in the ranking assignment and controlling for overall agreement in the

rating task. Hence location of the item in the set is most likely not a prime reason. On the 'why' of this finding we can only speculate. One plausible reason might be that when ranking work values is concerned it is acceptable to rank it among the top three items. After all, who does not work 'for a living'? In a rating task respondents might be more reluctant to rate its importance higher than other job values since it is less socially desirable. We have no means of controlling for socially desirable responding (in addition to overall agreement and primacy that are included in the model) with the current dataset. A second difference between ratings and rankings is observed in the case of "generous holidays". In the ranking data it is linked to extrinsic values, in the rating task both the extrinsic and the non-differentiation class assign similar importance to this issue. Regardless of these particularities, the contrast of the effect parameters of the extrinsic values on the second latent class compared with their estimated effect on the intrinsic latent class is pronounced. The reverse is true for the intrinsic items. This is highlighted in the two last columns in which we report the differences between the estimated effect parameter for the intrinsic latent class and the corresponding parameter for the extrinsic latent class. These contrast values are very similar across occasions (T1 and T2) and across measurements (ratings and rankings). Hence it is safe to conclude that these two classes have a distinct view on the intrinsic versus extrinsic work values inventory.

The remaining classes are difficult to compare across measurement methods since they are divergent. Part of this is – of course – by very nature of the measurement method itself. A non-differentiation class can only be observed in a rating assignment. In a ranking assignment the potential non-differentiators are forced to make their choices. How non-differentiators react to a ranking assignment is a key topic in the following section. The fourth latent class in the rating task can be labeled as a class of non-differentiators since its effect sizes are small or even not significantly different from average. Even the 'higher' positive scores observed such as for "pleasant people to work with" and "people treated

equally at the workplace" are still the lowest observed scores for these two very popular items in the rating assignment. Only the items "a useful job for society" and "family friendly" score relatively higher than in other classes but the small negative values observed indicate their low overall preference in all classes. A content label can be assigned to the third latent class in the rating and the ranking assignment but the label is different. For the ranking approach the items "a useful job for society", "meeting people", "people treated equally at the workplace" and "family friendly" are the most preferred by respondents belonging to this third class and therefore we called this class the social work values class. In the rating approach only the items "meeting people" and especially "people treated equally at the workplace" are more preferred by respondents belonging to the third class and therefore this class receives an adjusted label in which social is restricted to other people (not society or the own family). Within each measurement method the results from first and second measurement are highly similar suggesting that the classes found are not an artifact. When we present the findings on measurement invariance for the two subsamples that received the same method on both occasions we can further elaborate on the latter interpretation.

4.5.3 Two of a Kind: Similarities between Ranking and Rating Data in Classifications into Work Values Profiles

Using particular methods to model choice preferences in ranking and rating data revealed similar latent class profiles as far as intrinsic and extrinsic work values are concerned irrespective of whether rating or ranking questions were used. That was the key finding reported in the previous section. The question now is to what extent respondents will be classified in the same latent class when alternative measurement methods are used on two different occasions.

In our design we defined four subsamples. Two of these subsamples received the same measurement method at both occasions and two subsamples received different methods. The inclusion of two subsamples that were measured twice with the same instrument has a two-folded purpose. First, they serve as a kind of standard for the comparison of similarity in classification when different methods are used on both occasions. After all, even when the same measurement is used we can hardly expect perfect correspondence between two measures. Random error causes variation. Furthermore, in our dataset there is a time-lag of two months between first and second measurement. Although it is hard to imagine why 'true' work values orientations would change in a short framework of only two months we cannot exclude the possibility of a true change in orientation. Consequently, if we want to evaluate the consistency in classification across measurement methods we need to compare it with a cross-classification when the same measurement method is used. Second, the repeated measurement part of our experiment is used to research the level of measurement invariance across occasions. This will be the final topic presented in the next section of this research. In this section we compare cross-classifications from separate measurements.

In Table 4.4 we present the estimated values from four analyses in which the saved posterior membership probability scores from the first step of the analysis are the input; we used the three-step approach to adequately estimate associations between the two waves. A table with the estimated parameters is presented in appendix C. In this section we report the estimated values that indicate the cell percentages in the T1 by T2 table. We first elaborate on the two samples (Tables 4.4.1 and 4.4.2) that were administered the same measurement method. Columns in this case refer to the classification from the second measurement and rows to the classification from the first measurement. The two samples that changed measurement method (Tables 4.4.3 and 4.4.4) differ in the order of which each method was administered. To facilitate comparison, the row variable refers to the ranking assignment and

Table 4.4 Estimated cell % and residual % (= deviance from expected cell % with statistical independence) per test-condition T1 x T2

4.4.1 Ranking x Ranking (subsample 1)

T2	T1 RANKING						
RANKING	Intrinsic cell %	residual %	Extrinsic cell %	residual %	Social cell %	residual %	Column total %
Intrinsic	31.5	+18.8	0.1	-12.9	0.3	-5.9	31.9
Extrinsic	5.8	-11.6	36.7	+18.9	1.1	-7.3	43.6
Social	2.6	-7.2	4.0	-6.0	17.9	+13.2	24.5
Row total %	39.8		40.8		19.3		100.0

4.4.2 Rating x Rating (subsample 4)

T2	T1 RATING								
RATING	Intrinsic cell %	residual %	Extrinsic cell %	residual %	People cell %	residual %	Non-differentiation cell %	residual %	Column total %
Intrinsic	16.6	+10.6	0.6	-6.2	0.6	-5.9	4.4	+1.4	22.1
Extrinsic	0.7	-9.5	24.2	+12.6	11.2	-0.1	2.1	-3.0	38.2
People	9.1	+1.3	2.7	-6.1	16.4	+7.8	0.8	-3.0	29.0
Non-differentiation	0.4	-2.4	2.9	-0.3	1.3	-1.8	6.0	+4.6	10.6
Row total %	26.8		30.4		29.6		13.3		100.0

4.4.3 Ranking x Rating (Subsample 3)

T2	T1 RATING								
RANKING	Intrinsic cell %	residual %	Extrinsic cell %	residual %	People cell %	residual %	Non-differentiation cell %	residual %	Column total %
Intrinsic	18.7	+10.5	0.1	-10.1	8.7	-0.1	3.8	-0.3	31.3
Extrinsic	3.4	-8.1	23.0	+8.7	11.7	-0.7	5.9	+0.1	43.9
Social	4.1	-2.4	9.5	+1.4	7.7	+0.8	3.5	0.2	24.8
Row total %	26.2		32.6		28.1		13.1		100.0

4.4.4 Rating x Ranking (Subsample 2)

T1	T2 RATING								
RANKING	Intrinsic cell %	residual %	Extrinsic cell %	residual %	People cell %	residual %	Non-differentiation cell %	residual %	Column total %
Intrinsic	19.1	+9.9	2.2	-13.4	16.2	+5.3	2.2	-1.8	39.6
Extrinsic	1.6	-7.9	26.2	+10.0	8.2	-3.1	5.2	+1.1	41.2
Social	2.5	-2.0	10.9	+3.4	3.1	-2.1	2.6	+0.7	19.1
Row total %	23.2		39.3		27.5		10.0		100.0

the column variable to the rating task in both tables. Values presented in the tables are observed cell percentages and their residuals that indicate the deviance compared to the expected cell percentages with statistical independence. Our primary interest will be in the comparison of classification into the intrinsic and extrinsic work values class in each subsample.

The cross-classification of respondents in the same latent class across repeated measures (Tables 4.4.1 and 4.4.2) is extremely consistent. Row and column percentages slightly differ, which can be either due to small differences in measurement or due to change in time. The maximum cell percentage possible is thus limited to the smallest corresponding row or column percentage.

Subsample 1 who received the ranking assignment twice shows a cell percentage of 31.5% that is classified in the intrinsic class on both occasions. This value is almost the same as the corresponding row percentage of 31.9% and also close to the 39.8% column value. Similarly, the cell percentage of 36.7% in the extrinsic class on both waves is highly similar to the respective row 43.6% and column 40.8%. A similar observation is made in case of the third social latent class. The large positive residuals on the diagonal of consistent classifications confirm this interpretation.

The rating latent class model includes 4 latent classes. Here, cross-classification is a little bit more diffuse than in the case of the ranking assignment. The larger positive residuals on the diagonal of consistent classifications are still observed but are less pronounced in the case of the third “people” latent class and in particular when looking at the classification of the non-differentiators. The consistent scoring on the intrinsic and extrinsic latent class, however, is again clearly observed. 16.6% of respondents are classified as intrinsic on both occasions. This value needs to be compared to the 22.1 row percentage and the 26.8 column

percentage. Consistent scoring is similar in case of the extrinsic class with 24.2 cell percentage, 38.2 row percentage and 30.4 column percentage. Our overall interpretation of the results in Tables 4.4.1 and 4.4.2 is that when intrinsically work values oriented respondents and extrinsically oriented respondents answer to either ranking on rating questionnaires they tend to respond consistently across occasions. The intriguing next question is now: will they score consistently when the question format changes from first to second measurement?

The answer to the latter question is boldly: 'yes'. Whether the ranking questionnaire was administered after (Table 4.4.3) or before (Table 4.4.4) the rating questionnaire, in each case we observe positive residual values of consistent scoring in the case of the intrinsic latent class (+10.5% and +9.9%) and in the case of the extrinsic latent class (+8.7% and +10.0%). Comparing the cell percentages with column and row percentages is less straightforward than in Tables 4.4.1 and 4.4.2 since row and column distributions refer to two different measurement instruments (ratings and ratings) that have a different number of latent classes. Comparison of cell percentages with the column equivalence is a logical choice since the four latent classes model indicates the maximum percentage that might return in the table of cell percentages. This comparison reveals that the intrinsic cell percentages are closer to the column percentages than the extrinsic cell percentages are, compared to their column percentages. Hence, the consistency in classifying respondents in the intrinsic latent class across measurement method (rating versus ranking) is somewhat higher than in the case of the extrinsic latent class.

In the previous section we already argued that the third latent class in the ranking assignment and the third and fourth latent class in the rating assignment seem to be typical to each method separately. As such we did not expect particular relationships between them

when the measurement method changed from first to second measurement. This is confirmed by the results. We like to underscore the results regarding the fourth "non differentiators" latent class. One criticism to the use of ranking data is that respondents are arbitrarily forced to make choices and hence random choices might occur in case respondents do not make difference in assigning importance to the different items (Davis, Dowley, & Silver, 1999). Our results show that respondents that were classified as non-differentiators at first measurement (Table 4.4.3) contribute proportional to each of the latent classes of the ranking assignment at T2 since residual cell percentages are smaller than 1. Similarly, respondents classified in one of the three latent classes in the ranking assignment at T1 are proportionally allocated to the class of non-differentiators at T2. Hence, it is safe to conclude that 'forcing' non-differentiating respondents to make choices does not bias latent class identification in a ranking assignment.

4.6 Measurement Invariance of Measurement Methods

In the previous section we presented comparisons based on separate measures at first and second wave and we compared whether respondents were classified in similar latent classes across time. This was an obvious choice since we compared the groups who received the same measurement method twice with groups that differed in the measurement method used across time points. An important consequence of this procedure is that by definition there is no measurement invariance imposed upon the data, which is not possible if measurement methods are different. Recall that measurement invariance means that exactly the same measurement model applies in each measurement of a repeated measurement design. Hence the consistency in responding reported in the previous section is primarily conceptual. A more thorough test of consistency of the ranking and rating method is possible when on both

occasions use was made of the same measurement method. This is a significant addition to the previous analysis since if measurement invariance across modes is established we have evidence that the meaning assigned to the set of questions was indeed comparable across measurement occasions and that measurement equivalence is being achieved (Billiet & Davidov, 2008).

Testing for measurement invariance across occasions (Horn & McArdle, 1992) involves the formal comparison of models in which equality constraints across time points are compared with the estimates in which effects of the latent classes are freely estimated. Table 4.5 shows the goodness of fit measures of models with and without equality constraints.

In Table 4.5 we present model fit estimates of two models for each of the two measurement methods. The first model presented for the ranking and rating data has equal results as when one would perform the analyses on the two measurement occasions separately. This model does not impose equality restrictions on the intercept and slope values of the measurement model. Associations between latent variables across measurements are included. The second model does impose equality restriction on the intercept and slope estimates and in the case of the rating task also on the random intercept. Associations between latent variables are again estimated.

Based on the Bayesian Information Criterion (BIC) we conclude that the model with equality restrictions is preferred, indicating measurement invariance in both the ranking and rating task. AIC and AIC3 confirm the result in the case of ranking data but not when rating data is used. The difference between BIC and AIC(3) is that BIC not only penalizes for the number of parameters in the model but also for sample size. Nevertheless, the results indicate

Table 4.5 Model fit measurement equivalence analyses

Ranking data					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
Without equality restrictions	-31361	63509	62934	63040	106
With equality restrictions	-31382	**63187**	**62878**	**62935**	57
Rating data					
	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar
Without equality restrictions	-57405	116131	**115165**	**115343.3**	178
With equality restrictions	-57627	**115975**	115448	115545.4	97

Note: Values in bold indicate the smallest goodness of fit value

that the ranking results are to a large extent consistent across measurements occasion with the rating data producing a fair amount of consistency as confirmed with BIC.

In Appendix D we present the measurement model of the models with equality restrictions. Overall the results resemble the findings presented in Table 4.3. Of particular interest is the information regarding the associations between measurements at T1 and T2. Similar as in Table 4.4 we present the estimated cell percentages in the cross-classification of the measurement at T1 with T2 with the corresponding residual given statistical independence (see Table 4.6). As expected the associations are somewhat more articulated, but the resemblance with the results in Table 4.4 is close. This means that the observed associations reported in Table 4.4 are not an artifact of separate measurements at two occasions. As such we are confident that the associations presented in Table 4.4, when the measurement method was changed between T1 and T2, largely represent true associations.

Table 4.6 Estimated cell % and residual % (= deviance from expected cell % with statistical independence) per test-condition T1 x T2

4.6.1 Ranking

T2		T1 RANKING						
RANKING		Intrinsic cell %	residual %	Extrinsic cell %	residual %	Social cell %	residual %	Column total %
	Intrinsic	32.9	+21.5	1.0	-13.3	0.1	-8.3	33.9
	Extrinsic	0.3	-12.6	37.7	+21.5	0.6	-8.9	38.6
	Social	0.2	-9.0	3.4	-8.2	23.9	+17.1	27.5
	Row total %	33.4		42.1		24.5		100.0

4.6.2 Rating

T2		T1 RATING								
RATING		Intrinsic cell %	residual %	Extrinsic cell %	residual %	People cell %	residual %	Non-differentiation cell %	residual %	Column total %
	Intrinsic	17.6	+12.1	0.4	-8.4	2.6	-3.6	3.1	-0.1	23.7
	Extrinsic	0.9	-7.6	29.1	+15.6	3.7	-5.7	2.5	-2.3	36.2
	People	4.1	-2.7	5.6	-5.3	18.4	+10.9	1.1	-2.8	29.2
	Non-differentiation	0.8	-1.7	2.2	-1.8	1.3	-1.6	6.6	+5.1	10.9
	Row total %	23.4		37.3		26.0		13.3		100.0

4.7 Conclusion and Discussion

The key argument in this study is that there are segments within a population that respond similar to rating and ranking questions used to measure work values. To that purpose we investigated 1) whether the answers given by respondents at two measurement occasions are comparable irrespective of whether the respondents received a rating or a ranking measurement procedure and 2) how consistent these results were over time. A modified form-resistant hypothesis was adopted by arguing that it is important to take into account the format specific features of each measurement procedure which if not being controlled for can make it hard to match the results of different measurement methods. The method specific features controlled for in the current study are the primacy effect for the ranking data and the overall liking and non-differentiation for the rating data.

In searching for segments that reveal similar preferences in work values we needed to adopt a research approach that deviates from what has been used in previous research. First, instead of using a factor-analytic approach we used a latent class choice modeling approach which allowed us to distinguish between groups of respondents with similar response patterns in both the ranking and the rating method. These groups constitute homogeneous segments in the population that share a similar preference structure. Second, instead of adjusting the covariance structure of ranking data to eliminate the ipsativity of the data – a procedure suggested by Jackson and Alwin (1980) – we directly modeled the raw data in such a way that it reveals relative preferences. We also used a model that allowed to research relative preferences with rating data. This model implied the use of a random intercept to control for overall agreement. The measurement part of the model than also identifies relative preference structures similar to the model used with ranking data. The principal finding of this research is that respondents classified into either the intrinsic or extrinsic work values classes are

consistently classified as such across occasions even if the measurement method, i.e. ranking versus rating, changes in time. Other latent classes were method specific, a social work values class for the ranking assignment and a people oriented work values class and non-differentiating class for the rating assignment. These method specific classes were found consistently over the two measurement occasions when the same measurement method (rating or ranking) was used.

The within-subjects design thus enabled us to investigate how consistent the classifications were across measurement methods on two measurement occasions. We found it to be surprisingly high. Our modified form-resistant hypothesis stated that specific segments could be expected to emerge from either ranking or rating. We were particularly interested in finding out how non-differentiators in the rating assignment would respond to a ranking assignment in which they are forced to make a priority ranking of work items. The cross-classifications showed that non-differentiating respondents contributed proportionally to each of the latent classes in the ranking approach irrespective of whether they first rated and then ranked or vice versa. Thus, forcing non-differentiating respondents to choose does not lead to biases in the ranking results.

In the last part of this study, a more profound test of the consistency of receiving the same measurement method twice was conducted. We tested whether imposing the same measurement model across both measurement occasions would increase model fit. If so, we could conclude that measurement invariance is established and that the results are not an artifact of the method. The answer is confirmative: model fit improved and on top of that associations between repeated measures became more pronounced.

An inevitable limitation of this study was that we compared ratings and rankings in one particular context, namely work values. The question remains to what extent these

findings can be generalized to other types of concepts for which both the ranking and rating approach can be used. We also used a long items list which we thought would be most challenging in finding similarity in results. Whether similarity in results depends on the length of the items list remains to be researched. The methods used in this research, however, are also applicable with shorter lists of items. To applied researchers who consider using one or both of our methods, depending on whether they used ranking and/or rating scales we advise to develop a design that allows to control for method specific features such as primacy. We like to think of the method used in this research as being semi-exploratory. It is not completely exploratory since the research starts with a preconceived measurement model. In the ranking assignment, for instance, we included an effect of primacy and checked whether it improved measurement fit. This is typical to what is called confirmatory measurement modeling. At the same time, our models are exploratory in the sense that specific response pattern are revealed when adding latent classes to previous models. We regard this as a strength of our approach. Non-differentiating, for instance, was a response pattern that emerged from the data. We did not explicitly model it.

We hope that the current study has shown the usefulness of the latent class segmentation approach for both the comparison of rating and ranking data as for checking the consistency of the data over time. Using the approach in which we transformed ratings into relative preferences to compare this data with the ranking data we were able to show that rankings and ratings do produce results that are more similar than was previously assumed.

APPENDIX B: Examples of Latent GOLD Syntax

Latent Class Analysis Ranking Data

```
variables
caseidnomem_encr;
repscalesweight;
choicesetidversion_AB ;
dependent Choice ranking;
independent Rank123;
attribute _Constants_, Primacy ;
latent
Classnominal 3;
equations
Class<- 1 ;
Choice <- _Constants_|Class + Primacy Rank123 ;
```

Latent Class Analysis Rating Data

```
variables
caseidnomem_encr;
dependentRating;
independentItemnr nominal;
latent
CFactor1 continuous,
     Class nominal 4;
equations
 (1) CFactor1 ;
Class<- 1 ;
  Rating <- 1 | Class + CFactor1 + Itemnr | Class ;
```

Regressing T2 probabilities on T1 probabilities: step-3 proportional ML approach (example syntax Rank-Rate condition)

```
step3 proportional ml;
variables
// caseidnomem_encr;
latent
    Meting1 nominal posterior = ( Rank1.Class#1 Rank1.Class#2 Rank1.Class#3),
    Meting2 nominal posterior = ( Rate2.Class#1 Rate2.Class#4 Rate2.Class#2 Rate2.Class#3) ;
equations
    Meting1 <- 1;
    Meting2 <- 1 + Meting1;
```

APPENDIX C: Estimated effect parameters regressing T2- on T1- probabilities -

Results from the step-3 proportional ML approach

C3.1 Ranking x Ranking (subsample 1)

T2	T1 RANKING					
RANKING	Intrinsic beta	s.e.	Extrinsic Beta	s.e.	Social beta	s.e.
Intrinsic	2.545	(0.954)	-1.867	(1.468)	-0.678	(1.439)
Extrinsic	-0.902	(0.664)	1.999	(0.804)	-1.097	(0.985)
Social	-1.643	(0.672)	-0.133	(0.799)	1.775	(0.794)

C3.2 Rating x Rating (subsample 4)

T2	T1 RATING							
RATING	Intrinsic beta	s.e.	Extrinsic Beta	s.e.	People beta	s.e.	Non-differentiation beta	s.e.
Intrinsic	2.136	(0.410)	-1.477	(0.690)	-1.445	(0.620)	0.786	(0.306)
Extrinsic	-1.668	(0.630)	1.586	(0.314)	0.725	(0.301)	-0.643	(0.303)
People	0.889	(0.331)	-0.538	(0.331)	1.157	(0.256)	-1.508	(0.332)
Non-differentiation	-1.356	(0.728)	0.429	(0.373)	-0.437	(0.391)	1.365	(0.284)

C3.3 Ranking x Rating (Subsample 3)

T2	T1 RATING							
RANKING	Intrinsic beta	s.e.	Extrinsic Beta	s.e.	People beta	s.e.	Non-differentiation beta	s.e.
Intrinsic	1.704	(0.610)	-2.782	(1.716)	0.568	(0.598)	0.510	(0.615)
Extrinsic	-1.147	(0.432)	1.627	(0.868)	-0.281	(0.338)	-0.199	(0.356)
Social	-0.558	(0.417)	1.155	(0.876)	-0.287	(0.349)	-0.311	(0.378)

C3.4 Rating x Ranking (Subsample 2)

T1	T2 RATING							
RANKING	Intrinsic beta	s.e.	Extrinsic Beta	s.e.	People beta	s.e.	Non-differentiation beta	s.e.
Intrinsic	1.361	(0.387)	-1.509	(0.546)	0.637	(0.296)	-0.489	(0.428)
Extrinsic	-1.157	(0.523)	0.929	(0.317)	-0.098	(0.266)	0.327	(0.310)
Social	-0.204	(0.415)	0.581	(0.335)	-0.539	(0.322)	0.162	(0.339)

Appendix D: Parameter estimates of equivalent measurement models (ranking twice versus rating twice) with association between the latent classes at two time-points

	Ranking 3Class model			Rating 4Class model			
Items	Intrinsic	Extrinsic	Social	Intrinsic	Extrinsic	People	Non-diff
Meeting abilities	**2.111***	0.333*	0.678*	**1.395***	-0.117*	0.923*	0.286*
Responsible job	**0.550***	-0.910*	-0.722*	**0.325***	-1.419*	-0.428*	-0.299*
Interesting	**1.167***	0.301*	-0.405*	**1.273***	-0.426*	0.696*	0.155*
Achievesomething	**0.683***	-0.486*	-0.447*	**0.461***	-0.749*	-0.013	-0.092
Have a say	**-0.386***	-1.364*	-1.287*	**-0.132***	-1.253*	-0.786*	-0.257*
Use initiative	**0.112**	-0.356*	-0.228*	**0.662***	-0.278*	0.402*	0.019
Learn new skills	**-0.157***	-0.422*	-0.158	**0.361***	-0.457*	0.201*	0.106
Useful for society	0.011	-1.062*	**0.678***	-0.561*	-0.430*	-0.590*	**-0.325***
Meeting people	0.284*	-0.467*	**1.418***	0.057	0.168*	**0.368***	-0.170
People equally treated	-0.019	0.359*	**1.161***	0.777*	1.237*	**1.400***	0.405*
Family friendly	-1.606*	-1.027*	**-0.457***	-1.108*	-0.350*	-1.060*	**-0.104**
Holidays	-1.524*	**-0.777***	-1.249*	-1.207*	**-0.162***	-1.107*	-0.262*
Pleasantpeople	0.893*	**1.986***	1.731*	0.659*	**1.490***	1.414*	0.499*
Pay	1.010*	**2.365***	-0.269	-0.043	**0.736***	0.293*	0.209*
Job security	-0.530*	**0.845***	-0.493*	-0.677*	**0.985***	-0.104	0.014
Goodhours	-0.715*	**0.900***	0.425*	-0.582*	**0.766***	-0.14	0.118
No pressure	-1.884*	**-0.218***	-0.376*	-1.659*	**0.259***	-1.469*	-0.302*
Primacy effect		0.638*					
				Random Intercept	0.789*		

Notes: * parameter estimate is at least twice the standard error; Values in bold indicate for each item the highest preference value over all latent classes in each model

CHAPTER 5

Conclusion and Discussion

The purpose of the research that is reported in this dissertation was to answer the following research question: how comparable and consistent are measurements of personal values that result from the application of the rating and ranking approach when we account for the method-specific features of each response format? Stated much more simply, this question boils down to this: *are ratings and rankings actually two of a kind?* In previous studies comparing ratings and rankings mixed results were found with respect to the similarity of the two measurement methods (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; Maio, Roese, Seligman, & Katz, 1996; McCarty & Shrum, 2000; Ovadia, 2004). Krosnick and Alwin (1987, 1988) were the first to compare both measurement methods while accounting for method-specific features of each method. They came up with the "form-resistant correlation hypothesis" stating that observed correlations between values should not differ that much across different measurement methods. In the current study we used the form-resistant hypothesis in a less strict form, looking for similarities in preference structures between ratings and rankings while also allowing for inevitable differences between the two formats.

To answer the research question, we have developed a survey-experiment that enabled us not only to make a between-subjects comparison of the results of ratings and rankings, but also a within-subjects comparison. We investigated the rating-versus-ranking issue for the case of personal work values, which are often studied by substantive researchers and which

lend themselves very well to an operationalization in terms of both ratings and rankings. We have used state-of-the-art advanced statistical modeling techniques to assess the comparability and consistency of both measurement methods. The method used in the current study (latent class modeling approach) deviates from the method used in previous studies (factor analysis) comparing ratings and rankings. When comparing the results obtained with the ranking and the rating formats it is important to control for method-specific features which may have a biasing influence on the results, making the two approaches less comparable. The method-specific features that we controlled for in this study were response styles like the response order effect for the ranking format and the overall importance level and non-differentiation for the rating format.

We started with the introduction of a new approach to measure and control for response order effects. While in previous research it is assumed that a full randomization approach controls the existence of response order effects, the approach used here makes it possible to get an idea of the size of the response order effect being present and to statistically control for it in the measurement model even when only two different orders of the items were administrated. We found evidence that controlling for response order effects improved the fit of the model and that parameter estimates of items in this case shown first to respondents changed substantially. This study showed the necessity of controlling for response order effects if these effects do exist in the data, since conclusions can change substantially once the response order effect is being accounted for. The question remains whether response order effects are actually being ruled out by using full randomization of item order, but this issue can be investigated by using the approach described in this dissertation.

Next, we used the previously described approach to control for response order effects in ranking data when comparing these results with the rating approach. However, the rating

approach is not free of any response biases either, with the overall level of agreement being a potentially important source of bias. We accounted for the overall importance level by applying a model-based transformation of rating data into relative preferences, or, in other words, by 'ranking the ratings'. Also, when analyzing the rating data using this latent class segmentation approach it became clear that a group of non-differentiating respondents could be identified. By having these respondents in a separate latent class, this group no longer influenced the main results in which we were interested. The approach taken here, to change ratings into relative preferences instead of transforming the rankings, is different from the approach that has been used in previous research. The approach used in the current study makes it possible to use all information obtained by both the rating and the ranking task. Comparing the rating and ranking approach in a between-subjects study, we were able to find two latent segments with similar meanings and one group for which the interpretation was different, dependent on whether the ranking or rating approach was being used. This led us to conclude that accounting for the inevitable differences between the two methods (like for example caused by response style behavior) enables us to find groups of respondents with similar response patterns. The findings also showed that it is possible to use the rating approach instead of the more cumbersome ranking approach if a researcher is interested in the measurement of relative preferences. The results of the rating approach showed resemblances with the results of the ranking approach. Transforming ratings into relative preferences can give researchers more information about subjects of interest, especially when it is likely that respondents would rate all of the alternatives shown to them as important (to a greater or lesser extent). Researchers should be aware of the possibility to use ratings in such a manner.

Then the questions arose how consistent the results of the between-subjects analyses were, how consistent respondents were in showing a non-differentiation response style and what happens with these respondents when they were forced to choose between the items

using the ranking approach. A within-subjects design allowed us to investigate these questions. The findings demonstrated that respondents were consistently classified in latent segments with a similar label at the two measurement occasions. Irrespective of whether rankings or ratings were used to administer the importance of work values, the meaning of each latent class with similar label remained the same. Of course the consistency of cross-classifications between latent segments obtained from using the same measurement method twice showed the strongest cross-classifications compared to the situation in which respondents received a different measurement method at both time-points. The implementation of measurement invariance analyses for the respondents that received the same measurement method twice also provided evidence that forcing the parameter estimates to be the same for the two measurement occasions resulted in a better fitting model for both the ranking and the rating data. The improvement of fit for the ranking data was larger than for the rating data, meaning that the ranking approach was more consistent in measuring values than the rating approach. The third content group which was found to be different for rankings and ratings also remained different when performing within-subjects analyses. So, the conclusions of the between-subjects analyses with respect to the similarities and differences between rankings and ratings remain the same. Further, we found that the segment of non-differentiating respondents at the first measurement occasion was strongly associated with the group of non-differentiating respondents at the second measurement occasion. Apparently this non-differentiating behavior depends on personal behavior and is not a tendency which occurs occasionally. When respondents belonging to the non-differentiation segment were given the ranking response format, these respondents contribute proportionally to each of the latent segments. From this we infer that the non-differentiating respondents do not bias the results when forced to discriminate between items by using a ranking format.

Of course, further research is needed to see how the analysis approach used in this dissertation behaves when investigating other subjects for which both the ranking and rating approach can be used or when using other kind of respondents instead of the trained respondents used in this dissertation. The trained respondents used in the current study are part of a panel and are therefore used to participate in web surveys on a regular basis, which can lead to different results with respect to the presence of response biases in the data. Do the main results of this dissertation still hold when investigating other respondents and other contexts? It can also be interesting to use different rating scales and ranking response formats for the comparison of these two measurement methods, to indicate whether or not this may have an influence on the results. The choice of rating scale and ranking response format in this dissertation was based on the ones that have been used most often by applied researchers to investigate (work) values. However, it may be the case that the amount of response style behavior is dependent on these questionnaire characteristics. For example, does the number of non-differentiating respondents decrease when the rating scale gets larger and a respondent has thus more bullet points to choose from on the 'important' side of the scale (ranging from 'not at all important' to 'very important'), presumably leading to more differentiation in the answers? Another issue which may be explored is whether the length of the item list influences the similarity of the results. In the current study we used a long item list but the approach described in all chapters can also be applied to shorter item lists.

This dissertation has shown the usefulness of the latent class segmentation approach for comparing results obtained by ratings and rankings (both between- as within-subjects), as this statistical modeling approach can provide a common ground for comparison – i.e. the latent classes – even if measurements of the same theoretical concept have been acquired with different measurement procedures (as in our case measuring work values with either ratings of rankings). In addition, this study underscores the necessity to control for method-specific

biases to end up with results which are highly similar for the two measurement methods. The latent class approach was both suitable to estimate rankings and ratings as such (using the full information obtained by ratings and rankings) and suitable for measuring and controlling for response biases. It is important that researchers are aware of the response biases that may be influencing the results and that (when they find response biases to be present in their data) they check whether controlling for the bias is actually necessary to improve the fit of their statistical models and how much parameter estimates are affected by this control for biases. Also, we showed how the rating approach can be used in a way to gather another kind of information, namely relative preferences instead of absolute level of agreement.

So, in conclusion: *Are ratings and rankings actually two of a kind?* Taking stock of the results that this study has provided this question must be answered in the affirmative, albeit with a proviso: *Yes*, to a large degree they are two of a kind, *but* only when one takes the survey-methodological intricacies of each question format into account. This means that substantive researchers can be relatively assured when they seek to operationalize their theoretical concepts either as ratings or rankings: they are measuring the same concept to a highly similar degree – but not identical –, provided that they account for the specific confounding response style characteristics of each measurement approach.

References

Allison, P. D., & Christakis, N. A. (1994). Logit models for sets of ranked items. *Sociological Methodology, 24,* 199-228.

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49,* 535-552.

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*(2), 229-247.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology, 43*(1), 272-311.

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly, 18*(3):271-278.

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*(4), 542-562.

Böckenholt, U. (2002). Comparison and choice: Analyzing discrete preference data by latent class scaling models. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 163-182). Cambridge: Cambridge University Press.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis, 12*(1), 3-27.

Braithwaite, V. A., & Law, H. G. (1985). Structure of human values: Testing the adequacy of the Rokeach Value Survey. *Journal of Personality and Social Psychology, 49*(1), 250-263.

Campbell, D. T., & Mohr P. J. (1950). The effect of ordinal position upon responses to items in a check list. *Journal of Applied Psychology, 34*(1), 62-67.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292-303.

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research, 22*(2), 214-247.

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(1), 55-77.

Cheung, M. W. L. (2004). A direct estimation method on analyzing ipsative data with Chan and Bentler's (1993) method. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(2), 217-243.

Cheung, M. W. L. (2006). Recovering preipsative information from additive ipsatized data: A factor score approach. *Educational and Psychological Measurement, 66*(4), 565-588.

Croon, M. A. (1989). Latent class models for the analysis of rankings. In G. de Soete, H. Feger & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 99-121). Amsterdam: Elsevier Science Publishers.

Cunningham, W. H., Cunningham, I. C. M., & Green, R. T. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly, 41*(3), 379-384.

Davis, D. W., Dowley, K. M., & Silver, B. D. (1999). Postmaterialism in world societies: Is it really a value dimension? *American Journal of Political Science, 43*(3), 935-962.

DeCarlo, L. T., & Luthar, S. S. (2000). Analysis and class validation of a measure of parental values perceived by early adolescents: An application of a latent class model for rankings. *Educational and Psychological Measurement, 60*(4), 578-591.

De Chiusole, D., & Stefanutti, L. (2011). Rating, ranking , or both? A joint application of two probabilistic models for the measurement of values. *TPM - Testing, Psychometrics, Methodology in Applied Psychology, 18*(1), 49-60.

De Witte, H., Halman, L., & Gelissen, J. (2004). European work orientations at the end of the twentieth century. In W. Arts & L. Halman (Eds.), *European values at the turn of the millenium* (pp. 255-279). Leiden/Boston: Brill.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods, 17*(1), 30-52.

Duffy, R. D., & Sedlacek, W. E. (2007). The work values of first-year college students: Exploring group differences. *Career Development Quarterly, 55*(4), 359-364.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29*(1), 115-126.

Elizur, D. (1984). Facets of work values: A structural analysis of work outcomes. *Journal of Applied Psychology, 69*(3), 379-389.

Elizur, D. (1994). Gender and work values: A comparative analysis. *Journal of Social Psychology, 134*(2), 201-212.

Elizur, D., Borg, I., Hunt, R., & Beck, I. M. (1991). The structure of work values: A cross cultural comparison. *Journal of Organizational Behavior, 12*(1), 21-38.

Fuchs, M. (2005). Children and adolescents as respondents: Experiments on question order, response order, scale effects and the effect of numeric values associated with response options. *Journal of Official Statistics, 21*(4), 701-725.

Furnham, A., Petrides, K. V., Tsaousis, I., Pappas, K., & Garrod, D. (2005). A cross-cultural investigation into the relationships between personality traits and work values. *Journal of Psychology, 139*(1), 5-32.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72*(5), 892-913.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – A modified latent structure approach. *The American Journal of Sociology, 79*(5), 1179-1259.

Hagenaars, J. A. P. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Newbury Park, CA: Sage Publications.

Halman, L. (1996). *Individualization and the fragmentation of work values: Evidence from the European Values Study*. WORC-Paper 96.07.013. WORC - Work and Organization Research Centre. Tilburg.

Harzing, A. W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., Espejo, A., Ferreira, R., Giroud, A., Koester, K., Liang, Y. K., Mockaitis, A., Morley, M. J., Myloni, B., Odusanya, J. O. T., O'Sullivan, S. L., Palaniappan, A. K., Prochno, P., Choudhury, S. R., Saka-Helmhout, A., Siengthai, S., Viswat, L., Soydas, A. U., & Zander, L. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research?.*International Business Review, 18*(4), 417-432.

Horn, J. L., & McArdle J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144.

Inglehart, R. (1977). *The silent revolution: Changing values and political styles among Western publics*. Princeton, Guildford: Princeton University Press.

Inglehart, R. (1990). *Culture Shift in Advanced Industrial Society.* Princeton, N.J.: Princeton University Press.

Jackson, D. J., & Alwin, D. F. (1980). The factor analysis of ipsative measures. *Sociological Methods & Research, 9,* 218-238.

Jacoby, W. G. (2011). *Measuring value choices: Are rank orders valid indicators?* Paper presented at the 2011 Annual Meetings of the Midwest Political Science Association, Chicago, IL.

Kalleberg, A. L. (1977). Work values and job rewards: A theory of job satisfaction. *American Sociological Review, 42*(1), 124-143.

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for the external analysis of choice data. *International Journal of Research in Marketing, 11*(5), 451-464.

Kashefi, M. (2011). Structure and/or culture: Explaining racial differences in work values. *Journal of Black Studies, 42*(4), 638-664.

Klein, M., Dülmer, H., Ohr, D., Quandt, M., & Rosar, U. (2004). Response sets in the measurement of values: A comparison of rating and ranking procedures. *International Journal of Public Opinion Research, 16*(4), 474-483.

Kohn, M. L. (1969). *Class and conformity: A study in values*. Homewood, Ill.: The Dorsey Press.

Knoop, R. (1994). Work values and job satisfaction. *Journal of Psychology, 128*(6), 683-690.

Krosnick, J. A. (1992). The impact of cognitive sophistication and attitude importance on response-order and question-order effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 203-218). New York: Springer-Verlag.

Krosnick, J. A. (2000). The threat of satisficing in surveys: The shortcuts respondents take in answering questions. *Survey Methods Newsletter, 20*(1), 4-8.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*(4), 526-538.

Krosnick, J. A., & Presser S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263-313). Bingley: Emerald Group Publishing Limited.

Krosnick, J. A., & Schuman, H. (1988). Attitude intensity, importance, and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology, 54*(6), 940-952.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and prediction* (pp. 362-412). Princeton, N.J.: Princeton University Press.

Magidson, J., Eagle, T. C., & Vermunt, J. K. (2003). New developments in latent class choice models. In *Proceedings of the Sawtooth Software Conference, San Antonio, TX, April 15-17, 2003* (Vol. 10 pp. 89-112). Sequim, WA: Unknown Publisher.

Magidson, J., & Vermunt, J. K. (2006). Use of latent class regression models with a random intercept to remove the effects of the overall response rating level. In A. Rizzi & M. Vichi (Eds.), *COMPSTAT 2006: Proceedings in Computational Statistics* (pp. 351-360). Heidelberg: Springer.

Maio, G. R., Roese, N. J., Seligman, C., & Katz, A. (1996). Rankings, ratings, and the measurement of values: Evidence for the superior validity of ratings. *Basic and Applied Social Psychology, 18*(2), 171-181.

McCarty, J. A., & Shrum, L. J. (1997). Measuring the importance of positive constructs: A test of alternative rating procedures. *Marketing Letters, 8*(2), 239-250.

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly, 64*(3), 271-298.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly, 67*(1), 205-211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research, 20*(1), 60-103.

McFadden, D. (1986). The choice theory approach to marketing research. *Marketing Science, 5*(4), 275-297.

McFadden, D., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics, 15*(5), 447-470.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Miethe, T. D. (1985). The validity and reliability of value measurements. *The Journal of Psychology, 119*(5), 441-453.

Moore, M. (1975). Rating versus ranking in the Rokeach Value Survey: An Israeli comparison. *European Journal of Social Psychology, 5*(3), 405-408.

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research, 22*(1), 93-119.

Moors, G., & Vermunt, J. (2007). Heterogeneity in post-materialist value priorities. Evidence from a latent class discrete choice approach. *European Sociological Review, 23*(5), 631-648.

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research, 16*(1), 48-52.

Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology, 7*(5), 403-414.

Parsons, T., & Shils, E. A. (Eds.). (1962). *Toward a general theory of action*. New York: Harper.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, California: Academic Press, Inc.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

Rankin, W. L., & Grube, J. W. (1980). A comparison of ranking and rating procedures for values system measurement. *European Journal of Social Psychology, 10*, 233-246.

Ravlin, E. C., & Meglino, B. M. (1987). Effect of values on perception and decision making: A study of alternative work value measures. *Journal of Applied Psychology, 72*(4), 666-673.

Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. *Journal of Marketing Research, 17*(4), 531-536.

Rokeach, M. (1973). *The nature of human values*. New York: The Free Press.

Ros, M., Schwartz, S. H., & Surkiss, S. (1999). Basic individual values, work values, and the meaning of work. *Applied Psychology: An International Review, 48*(1), 49-71.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80*(1), 1-28.

Sacchi, S. (1998). The dimensionality of postmaterialism: An application of factor analysis to ranked preference data. *European Sociological Review, 14*(2), 151-175.

Scherpenzeel, A. C., & Das, M. (2010). "True" longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 77-104). Boca Raton: Taylor & Francis.

Schuman, H., & Presser S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, California: Sage Publications.

Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology, 53*(3), 550-562.

Siminski, P. (2008). Order effects in batteries of questions. *Quality & Quantity, 42*(4), 477-490.

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods, 1*(3), 121-138.

Super, D. E. (1962). The structure of work values in relation to status, achievement, interests, and adjustment. *Journal of Applied Psychology, 46*(4), 231-239.

Van Herk, H., & Van de Velden, M. (2007). Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality and Preference, 18*(8), 1096-1105.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450-469.

Vermunt, J. K., & Magidson, J. (2005a). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In L. A. van der Ark, M. A. Croon & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 41-62). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vermunt, J. K., & Magidson, J. (2005b). *Latent GOLD Choice 4.0 user's guide*. Belmont: Statistical Innovations, Inc.

Vermunt, J. K., & Magidson, J. (2013). *Latent Gold 5.0 upgrade manual*. Belmont, MA: Statistical Innovations Inc.

Vriens, I., Moors, G., Gelissen, J., & Vemunt, J. K. (in press). Controlling for response order effects in ranking items using latent choice factor modeling. *Sociological Methods & Research.*

Wollack, S., Goodale, J. G., Wijting, J. P., & Smith, P. C. (1971). Development of the survey of work values. *Journal of Applied Psychology, 55*(4), 331-338.

Summary

After years of discussing the appropriateness of ratings versus rankings for the measurement of values, still no consensus has been reached. The discussion stems from uncertainty whether values are hierarchically ordered in the human brain (in which case the ranking procedure would be the most appropriate method to use) or whether an individual evaluates each value independently of the evaluation of other values (which can be measured by using the rating method). Ideally, the results obtained by ratings and rankings should not be that different, since the only thing that differs between the two methods is the answering task that respondents receive. However, in previous research the results were mixed with respect to the comparability of the two measurement procedures. In this dissertation the results obtained by ratings and rankings are compared by using a different approach than the analysis procedure used in previous studies. We used a latent class modeling approach which made it possible to use ranking and rating data as such, while we at the same time accounted for method-specific features (like response biases) of each response format which may be biasing the results.

We started our research with the development of a new way to statistically control for response order effects. Instead of needing a full randomization of the order of all alternatives shown to respondents to eliminate the response order effect, it was shown that it is possible to measure and control for response order effects by having only two different orderings of the items. Comparing the results of the ranking data with and without controlling for response

order effects showed that the preference order of the items changed and therefore that the conclusions changed. It is thus important to check whether response order effects may be influencing the data and, if response order effects are present in the data, to control for these effects.

We continued our investigation by comparing the ranking method in which we controlled for the response order effect with the rating approach in which we controlled for the overall agreement tendencies of respondents (by transforming the rating data into relative preferences) and the level of non-differentiation. Based on the between-subjects comparisons it was shown that using the latent class segmentation method and controlling for the response biases described before, groups of respondents could be found with similar response patterns irrespective of the method being used (rating versus ranking). However, not all groups found present in the data had similar meanings. One of the three content groups was found to be divergent for the two approaches and thus for this group it did matter which approach was being used. So, allowing for differences between the two methods when comparing them makes it possible to gain results that are similar but it does not mean that all conclusions drawn are the same. Another interesting finding was that by transforming the rating data into relative preferences, we were able to measure 'relative preferences' classifications that were similar to the results of a ranking assignment for two out of the three content groups distinguished in our research.

The next step was to compare the stability of the latent groups found to be present in the ranking and rating data by using a within-subjects design with at least two months in between the two measurements. All groups recognized in the between-subjects analyses were found consistently over time, also the group that differed in meaning between the ranking and rating data. Measurement invariance was being tested for respondents that received the same measurement method twice and this resulted in a better model fit for the models in which the

parameter estimates were set to be equal between the two measurement occasions. We were also interested in what happened when respondents received a different measurement approach at both measurement occasions. The two groups with similar meanings between ratings and rankings showed consistent cross-classifications with their counterparts when another measurement method was being used, while the content group which differed in meaning between the two approaches did not show a particular relationship between the rating versus the ranking approach. Respondents belonging to the non-differentiation group were found to be contributing proportionally to each of the groups when these respondents received the ranking approach.

To summarize, we showed that by accounting for method-specific features of each response format and by using a latent class modeling approach, the results obtained by the ranking and the rating approach are more similar than previously assumed. We found evidence that it is important to check after data collection which response biases may be influencing your results and to control for these biases in the process of analyzing the data (not only when comparing ratings and rankings). Another subject of interest was the group of non-differentiating respondents which we found was stable over time (containing around 10% of the respondents) and thus seemed to be a personal answering style. When the non-differentiating respondents received the ranking approach they were evenly distributed over the distinguished groups, meaning that these respondents did not have a biasing influence on the results when the ranking procedure was being used.

Samenvatting (Summary in Dutch)

Na jaren van discussie over de geschiktheid van ratingschalen versus rankingschalen voor het meten van waarden is er nog steeds geen consensus bereikt over welke methode beter is. De discussie komt voort uit onzekerheid of waarden hiërarchisch geordend zijn in het menselijke brein (in dit geval zou de ranking-procedure de meest gepaste methode zijn om te gebruiken) of dat een individu elke waarde onafhankelijk van de evaluatie van andere waarden evalueert (hetgeen gemeten kan worden door het gebruik van de rating-methode). Idealiter zouden de resultaten verkregen met ratingschalen en rankingschalen niet erg verschillend moeten zijn, aangezien het enige verschil tussen de twee methodes de manier is waarop respondenten antwoord moeten geven op de vraag. Echter, in voorgaand onderzoek zijn de resultaten gemengd met betrekking tot de vergelijkbaarheid van de twee meetmethodes. In deze dissertatie vergelijken we de resultaten verkregen met ratingschalen en rankingschalen door het gebruik van een andere aanpak dan de analyse procedure gebruikt in voorgaande studies. Wij hebben een latente klassen aanpak gebruikt welke het mogelijk maakte om ranking en rating data als zodanig te modelleren, terwijl we tegelijkertijd rekening hebben gehouden met methode-specifieke kenmerken (zoals antwoordmeetfouten) van elke antwoordmethode die kunnen leiden tot vertekende resultaten.

We zijn ons onderzoek gestart met het ontwikkelen van een nieuwe manier om statistisch te controleren voor volgorde-effecten. Het is aangetoond dat in plaats van het nodig

hebben van een volledige randomisatie van de volgorde van alle alternatieven die respondenten te zien krijgen om volgorde-effecten te kunnen elimineren. Men kan volstaan met slechts twee verschillende volgordes van de alternatieven om volgorde-effecten te meten en hiervoor te controleren. Het vergelijken van de resultaten van de ranking data met en zonder controle voor volgorde-effecten heeft laten zien dat de preferentie-volgorde van de alternatieven veranderde en dat daardoor ook de conclusies veranderden. Het is dus belangrijk om te controleren of volgorde-effecten een invloed hebben op de data en wanneer volgorde-effecten aanwezig zijn in de data, om te controleren voor deze effecten.

We vervolgden onze zoektocht door het vergelijken van de ranking-methode waarbij gecontroleerd werd voor volgorde-effecten met de rating aanpak waarbij we controleerden voor het fenomeen dat respondenten (bijna) alle alternatieven belangrijk vinden (door het transformeren van de rating data in relatieve preferenties) en voor het niveau van niet-differentiatie. Gebaseerd op de vergelijkingen tussen respondentgroepen is aangetoond dat het gebruiken van de latente klassen segmentatiemethode en het controleren voor de antwoordmeetfouten zoals hiervoor beschreven ertoe hebben geleid dat groepen van respondenten gevonden zijn met vergelijkbare antwoordpatronen, ongeacht de gebruikte methode (rating versus ranking). Echter, niet alle gevonden groepen hadden een vergelijkbare betekenis. Eén van de drie inhoudelijke groepen week in betekenis af bij het vergelijken van de twee methodes en voor deze groep maakte het wel uit welke methode gebruikt was. Het toestaan van verschillen tussen de twee methodes bij het maken van een vergelijking maakt het mogelijk om resultaten te krijgen die meer overeenkomen, maar dat betekent niet dat alle getrokken conclusies per definitie hetzelfde zullen zijn. Een andere interessante bevinding was dat door het transformeren van de rating data in relatieve preferenties, we ‘relatieve preferentie’-classificaties konden meten die vergelijkbaar waren met de resultaten van de

ranking-opdracht voor twee van de drie in ons onderzoek onderscheiden inhoudelijke groepen.

De volgende stap was om de stabiliteit van de gevonden latente groepen in de ranking en rating data te vergelijken door het gebruik van een design waarbij herhaalde metingen van dezelfde respondenten met elkaar vergeleken zijn met een minimale periode van twee maanden tussen de twee metingen in. Alle erkende groepen uit de analyses tussen personen werden consistent over tijd bevonden, ook de groep die verschilde in betekenis afhankelijk van of de ranking- of rating-methode gebruikt was. Meetinvariantie werd getest voor respondenten die dezelfde meetmethode twee keer ontvingen en dit resulteerde in een betere modelkwaliteit voor modellen waarbij de parameterschattingen gelijk werden gezet voor de twee meetmomenten. We waren ook geïnteresseerd in wat er zou gebeuren wanneer een respondent een verschillend meetinstrument ontving op beide meetmomenten. De twee groepen met vergelijkbare betekenis tussen de ratingschalen en rankingschalen liet een consistente cross-classificatie zien met zijn tegenhanger wanneer een andere meetmethode was gebruikt, terwijl de groep die verschilde in betekenis tussen de twee methodes geen specifieke relatie liet zien tussen de rating- versus de ranking-aanpak. Respondenten die tot de niet-differentiërende groep behoorden waren evenredig verdeeld over elk van de groepen wanneer deze respondenten de ranking-methode ontvingen.

Samenvattend, we hebben aangetoond dat door rekening te houden met methode-specifieke kenmerken van elke antwoordmethode en door het gebruik van de latent klassen methode, resultaten verkregen door de ranking- en de rating-methode meer gelijk zijn aan elkaar dan voorheen werd aangenomen. Tevens hebben we bewijs gevonden dat het belangrijk is om na dataverzameling te controleren welke antwoordmeetfouten een invloed kunnen hebben op je resultaten en om voor deze fouten te controleren tijdens het analyseren van de data (niet alleen wanneer men ratingschalen met rankingschalen vergelijkt). Een ander

interessant onderwerp was de groep van niet-differentiërende respondenten welke in ons onderzoek stabiel was over tijd (bestaande uit ongeveer 10% van de respondenten) en daardoor lijkt dit een persoonlijke antwoordstijl te zijn. Wanneer niet-differentiërende respondenten de ranking-methode aangeboden kregen verspreidden zij zich gelijkmatig over de onderscheiden groepen, wat betekent dat deze respondenten geen vertekende invloed hebben op de resultaten verkregen met de ranking-methode.

Dankwoord

Mijn proefschrift was nooit zo geworden als dat het nu is zonder de hulp van mijn (co)promotoren, John Gelissen, Guy Moors en Jeroen Vermunt. Ik wil jullie bedanken voor het bieden van de mogelijkheid om dit proefschrift te schrijven en voor de begeleiding. Ik heb veel van jullie geleerd met betrekking tot het hele proces van het opzetten van een experiment, bedenken van een design waarmee verschillende onderzoeksvragen beantwoord kunnen worden, tot aan het schrijven van een (ook voor toegepaste onderzoekers) begrijpelijk artikel. Ook al was ik met mijn psychologenachtergrond misschien een beetje een vreemde eend in de bijt tussen de sociologen, toch ben ik van mening dat we er samen toch iets heel moois van hebben weten te maken.

Daarnaast dank ik CentERdata voor het verzamelen van de data die we voor alle hoofdstukken in dit proefstuk hebben kunnen gebruiken. Zonder de hulp van CentERdata was het moeilijk geweest om ons complexe design op een nette manier geïmplementeerd te krijgen. Ook wil ik graag het IOPS bedanken voor het bieden van de mogelijkheid om aanvullende cursussen te volgen en voor de interessante (en gezellige) congressen. Tevens wil ik alle MTO-collega's bedanken. Speciale dank gaat uit naar alle kamergenoten die ik in de vier jaar dat ik bij het MTO departement werkzaam was 'versleten' heb. Bedankt voor de gezelligheid, afleiding en hulp waar nodig. Mijn nieuwe collega's van het ABS/MEP-team wil ik bedanken voor het feit dat ze me met veel warmte in hun team hebben opgenomen.

Als laatste wil ik mijn familie en vrienden bedanken. Het is fijn om een goede en stabiele basis te hebben. Waar de meesten vooral hebben gezorgd voor een luisterend oor en de nodige afleiding in de afgelopen vier jaren, zijn er een aantal die ik apart wil bedanken vanwege hun directe bijdrage aan dit proefschrift of aan de verdediging. Robert, bedankt dat je mijn kaft hebt willen ontwerpen. Marianne en Susanne, heel blij ben ik dat jullie als paranimfen aan mijn zij zullen staan. Dirk, dank voor je hulp met de eerste formules en voor alle steun tijdens de afgelopen jaren.